Michał Sebastian Wróbel

Zastosowanie neuronowych systemów rozmytych w chemii

Promotor pracy
Prof. dr hab. Beata Walczak
Zakład Chemii Analitycznej
Instytut Chemii
Uniwersytet Śląski
w Katowicach

Instytut Chemii
Uniwersytet Śląski
Katowice, 2011

Składam serdeczne podziękowania
Pani prof. dr hab. Beacie Walczak
za opiekę naukową, poświęcony czas i cierpliwość,
a także Ivanie oraz Michałowi
za to, iż zawsze służyli mi dobrą radą.
W mojej pamięci na trwałe pozostaną osoby,
które towarzyszyły mi w niełatwej drodze do „dzisiaj".

ukochanemu Tacie

Spis treści

1 Wykaz skrótów

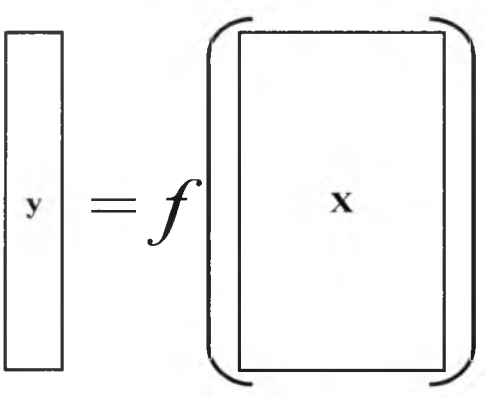
ANN – *artificial neural network* – sztuczna sieć neuronowa;
CART – *classification and regression trees* – drzewa klasyfikacji i regresji;
CCR – *correct classification rate* – procent poprawnie sklasyfikowanych obiektów;
DPLS – *discriminate partial least squares* – dyskryminacyjna metoda częściowych najmniejszych kwadratów;
DU – skrót od algorytmu Duplex, metody podziału obiektów na zbiory;
FCM – *fuzzy C-means clustering* – rozmyta metoda grupowania C-środków;
FIS – *fuzzy inference system* – system wnioskowania rozmytego;
GMP – *generalized modus ponens* – uogólniona reguła wnioskowania *modus ponens*;
KS – skrót od algorytmu Kennarda i Stone'a, metody podziału obiektów na zbiory;
NFS – *neuro fuzzy system* – system wnioskowania rozmytego;
NIR – *near infrared spectroscopy* – spektroskopia w bliskiej podczerwieni;
NMR – *nuclear magnetic resonance* – jądrowy rezonans magnetyczny;
PCA – *principal component analysis* – analiza czynników głównych;
PCs – *principal components* – czynniki główne;
PLS – *partial least squares* – metoda częściowych najmniejszych kwadratów;
RMSEP – *root mean square error of prediction* – pierwiastek średniego błędu kwadratowego przewidywania;
rPCA – *robust principal component analysis* – elastyczna analiza czynników głównych;
$\mathbf{X}$ – macierz;
$\mathbf{x}$ – wektor;
x – skalar;
$\boldsymbol{X}_{ml}, \boldsymbol{y}_{ml}$ – zbiór modelowy;
$\boldsymbol{X}_{mr}, \boldsymbol{y}_{mr}$ – zbiór monitoringowy;
$\boldsymbol{X}_{tt}, \boldsymbol{y}_{tt}$ – zbiór testowy;
$\mu_A(x)$ – funkcja aktywacji lub przynależności obiektu x do zbioru A;
ZM – zmienne wybrane metodą CART

2 Wstęp

Głównym problemem współczesnej chemii analitycznej jest jakościowe oraz ilościowe oznaczanie składu chemicznego różnorodnych próbek. Szczególnej uwagi wymagają próbki pochodzenia naturalnego, które odznaczają się złożonym składem chemicznym.

Przykładem analizy próbek naturalnych może być kontrola jakości oraz autentyczności produktów żywnościowych. Monitorowanie produktów spożywczych jest szczególnie ważne dla Unii Europejskiej. Działania unijne obejmują także ochronę konsumentów i producentów przed nieuczciwą konkurencją, np. przed rozmaitymi formami fałszowania produktów. Przykładami badań realizowanych w ramach unijnych projektów są np. program TRACE [1] oraz *Wine Data Base* [2]. Ich celem było potwierdzenie autentyczności i pochodzenia geograficznego wybranych produktów żywnościowych takich jak miód, woda mineralna, szynka, oliwa z oliwek, salami, kukurydza czy wino.

Modelowanie pochodzenia geograficznego czy autentyczności danego produktu opiera się na założeniu, iż informacja ta jest zawarta w jego unikalnym składzie chemicznym. Ten sam typ miodu, ale pochodzący z różnych obszarów danego kraju czy Europy, będzie różnił się zawartością makro- i mikroelementów. Szeroko rozwinięte współczesne techniki instrumentalne są źródłem bogatych informacji o analizowanych próbkach. Otrzymane sygnały instrumentalne nie rzadko mają długość tysięcy, a nawet setek tysięcy punktów pomiarowych. Ekstrakcję informacji z takich wielowymiarowych danych oraz ich analizę umożliwiają metody chemometryczne.



Rys. 1 Schemat wieloparametrowego modelu, gdzie **y** to zmienna zależna, a **X** to macierz zmiennych niezależnych

Modelując np. pochodzenie geograficzne danego produktu żywnościowego konstruuje się odpowiedni model matematyczny. Ogólne zależności pomiędzy modelowanymi danymi **X**, a modelowaną właściwością **y** przedstawia rysunek 1.

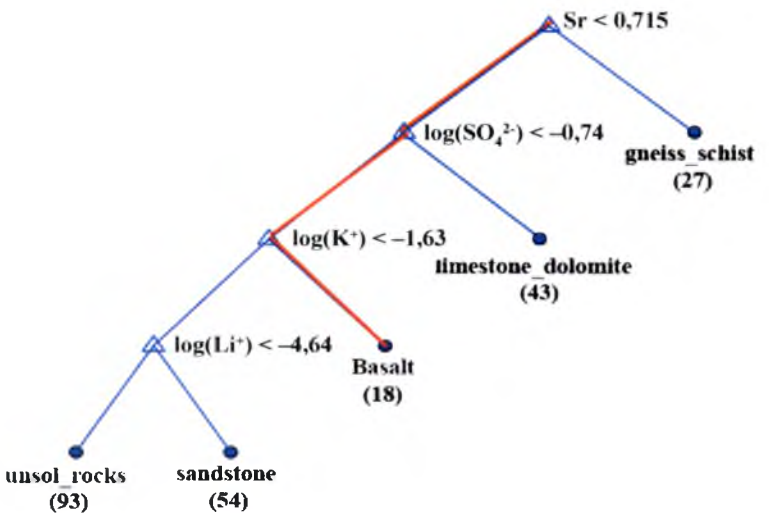
Chemiczne informacje o badanych próbkach zestawione są w formie macierzy **X**, której wiersze mogą zawierać np. widma w bliskiej podczerwieni (NIR) chromatogramy czy widma NMR. Macierz zmiennych niezależnych **X**, może być także tablicą zawierającą stężenia określonych jonów. Kodowanie kraju pochodzenia czy regionu jest dyskretne i najczęściej odbywa się w systemie binarnym lub bipolarnym. Oznacza to, iż np. próbkom wina pochodzącym z Republiki Czeskiej przyporządkowana zostaje wartość odpowiednio „0" lub „-1", natomiast próbkom pochodzącym z Węgier wartość „1". Mówi się wtedy o dyskryminacji. Zmienna zależna, **y**, może również być zmienną ciągłą i zawierać informacje np. o zawartości danej substancji w badanych próbkach. W takim przypadku mówi się o kalibracji.

Niejednokrotnie nie poprzestaje się na jednej metodzie modelowania danych. Celem jest znalezienie najlepszej metody, a więc takiej, która umożliwiłaby konstrukcję modelu obarczonego najmniejszym błędem przewidywania dla próbek z niezależnego zbioru testowego. Pomocna jest w tym znajomość zarówno wad i zalet różnych technik modelowania danych, a także wiedza o naturze modelowanego zjawiska lub procesu.

We wspomnianych wcześniej projektach TRACE oraz *Wine Data Base* wykorzystane zostały takie powszechnie stosowane metody modelowania danych jak: regresja czynników głównych (PCR) [3], metoda częściowych najmniejszych kwadratów (PLS) [3] oraz drzewa klasyfikacji i regresji (CART) [4]. Co prawda metoda CART ustępuje pozostałym wspomnianym technikom modelowania danych pod względem mocy predykcyjnej, jednakże pozwala na konstrukcję tzw. reguł logicznych. Reguły takie są pomocne przy interpretacji modelu. Ich zalety dostrzec można szczególnie wtedy, gdy zadanie interpretacji modelu powierza się osobom nieposiadającym specjalistycznej wiedzy, czy doświadczenia w danej dziedzinie.

3 Drzewa klasyfikacji i regresji

Drzewa klasyfikacji i regresji (z ang. *classification and regression trees*, CART) to technika modelowania danych znajdująca zastosowanie zarówno do problemów klasyfikacyjnych jak i kalibracyjnych [4]. Celem konstrukcji modelu CART jest podział wszystkich obiektów ze zbioru danych na jak najbardziej homogeniczne grupy, czyli zawierające obiekty najbardziej podobne do siebie. Otrzymany model można graficznie przedstawić w postaci binarnego drzewa decyzyjnego (Rys. 2).



Rys. 2 Drzewo modelu CART skonstruowane dla celów klasyfikacji próbek wody mineralnej w Europie pod względem ich pochodzenia geograficznego W ramach programu TRACE [1]

Takie drzewo to szereg parametrów i ich wartości znajdujących się w węzłach, pozwalających na grupowanie obiektów ze zbioru danych. Przykładowe drzewo widnieje również na rysunku 3b. W drzewie decyzyjnym można wyróżnić tzw. węzły „rodzice" i węzły „dzieci". Pierwszy od góry węzeł jest tylko rodzicem, a ostatnie węzły terminalne tylko dziećmi. Reszta węzłów pełni rolę zarówno dzieci (dla węzłów powyżej) jak i rodziców (dla węzłów poniżej).

3.1 Kryterium wyboru zmiennych

Kolejne węzły drzewa decyzyjnego opisują sposób podziału obiektów na coraz to bardziej homogeniczne grupy. Model wykorzystuje do tego celu parametry i ich wartości, biorąc pod uwagę zmienną zależną **y**. Drzewo przedstawia więc zbiór reguł logicznych opisujących przynależność próbek do konkretnych grup. Utworzone grupy obiektów należą do wzajemnie wykluczających się podprzestrzeni w przestrzeni modelowanych danych. Podprzestrzenie te oddzielone są hiperpłaszczyznami.

Konstruując drzewo klasyfikacji i regresji postępuje się według następującej procedury:

- konstruuje się drzewo o maksymalnej liczbie węzłów terminalnych;
- redukuje się ilość węzłów w drzewie, tzw. przycinanie drzewa (z ang. *tree pruning*);
- określa się optymalną kompleksowość drzewa (ilość węzłów terminalnych).

Obiekty do podzbiorów przypisywane są na drodze podziału rekurencyjnego. Jakość podziału opisuje funkcja zanieczyszczenia węzła, która dana jest następującym wzorem:

$$\Delta I(v,t) = I(t) - p_L I(t_L) - p_R I(t_R) \qquad 1$$

gdzie: $I(t)$ to zanieczyszczenie węzła t, v to parametr w danych **X**, dla którego dokonuje się podziału, natomiast proporcje podziału obiektów do prawego i lewego węzła dziecka opisują p_L i p_R.

Wartość parametru opisującego zanieczyszczenie danego węzła maleje wraz ze wzrostem homogeniczności (jednorodności) grupy obiektów należących do tego węzła. Jednorodność węzła opisują różne miary. Najbardziej popularną miarą jednorodności węzła jest entropia:

$$I(t) = -\sum_{i=1}^{k} p_i(t) ln[p_i(t)] \qquad 2$$

gdzie: k to numer grupy, p_i to proporcja obiektów z i-tej klasy w węźle t.

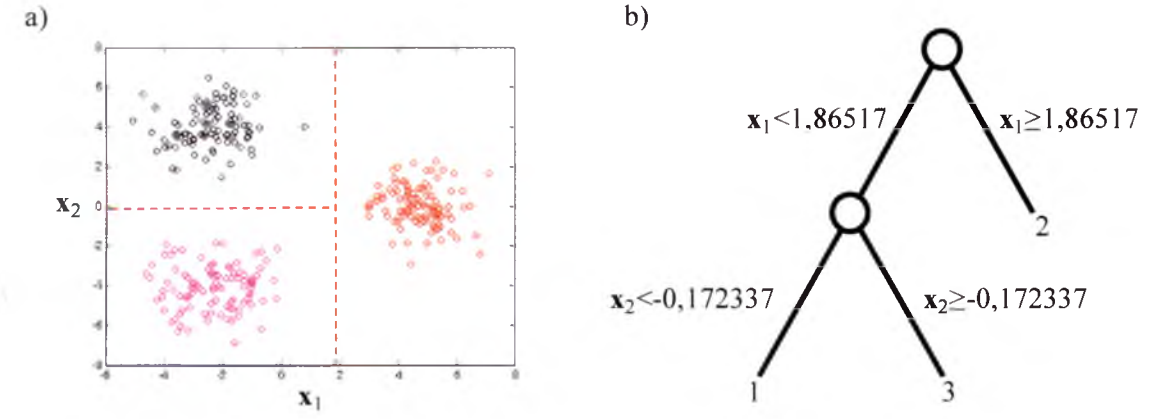
3.2 Optymalizacja i walidacja drzewa

Po skonstruowaniu drzewa o maksymalnej liczbie węzłów terminalnych przystępuje się do przycinania drzewa poprzez redukcję węzłów terminalnych. Celem przycinania drzewa jest uzyskanie modelu o optymalnej strukturze. Za optymalną strukturę drzewa uważa się strukturę z minimalną ilość węzłów, która pozwala na konstrukcje modelu przewidującego przynależność próbek z niezależnego zbioru testowego do grup z jak najmniejszym błędem. Optymalna struktura drzewa jest charakteryzowana przez kryterium kosztu-złożoności $R_\alpha(T)$:

$$R_\alpha(T) = R(T) + \alpha|T| \qquad 3$$

gdzie: R(T) to błąd klasyfikacji modelu o danej strukturze, |T| to ilość węzłów terminalnych a α to parametr z przedziału {0,1}. Jeśli dla kilku różnych drzew wartość parametru $R_\alpha(T)$ jest równa, wybiera się to drzewo, dla którego parametr α przyjmuje najmniejsze wartości. Optymalną strukturę drzewa określa się w oparciu o tzw. zbiór monitoringowy lub na drodze walidacji krzyżowej.

Na poniższym rysunku przedstawiono przykładowe dane w dwuwymiarowej przestrzeni pomiarowej oraz skonstruowane dla tych danych drzewo CART. Przerywane linie przedstawiają granice podziału przestrzeni danych utworzone przez model CART.

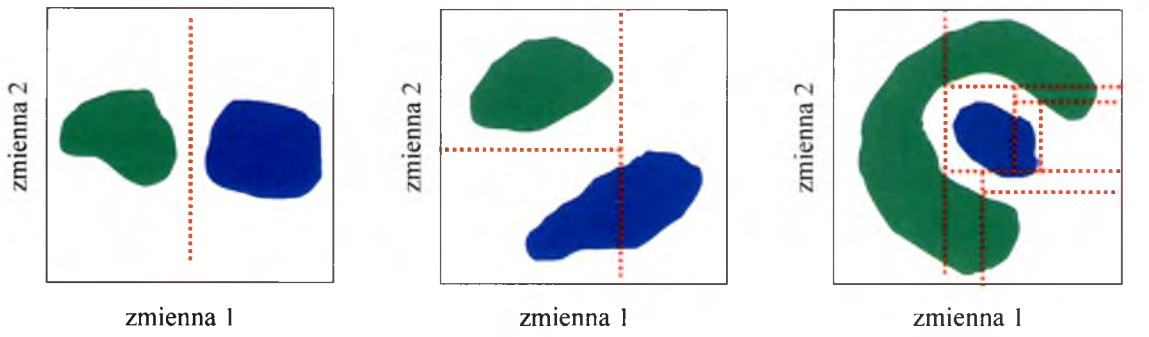


Rys. 3 a) Symulowane dwuwymiarowe dane zawierające trzy grupy wraz z podziałami utworzonymi przez model CART (zaznaczonymi linią przerywaną); b) Drzewo decyzyjne CART odpowiadające prezentowanym obok danym

Po skonstruowaniu modelu, można przewidywać przynależność nowych próbek do jednej z grup obiektów. Dokonuje się tego w oparciu o wartości parametrów wyznaczonych jako decyzyjne przez metodę CART.

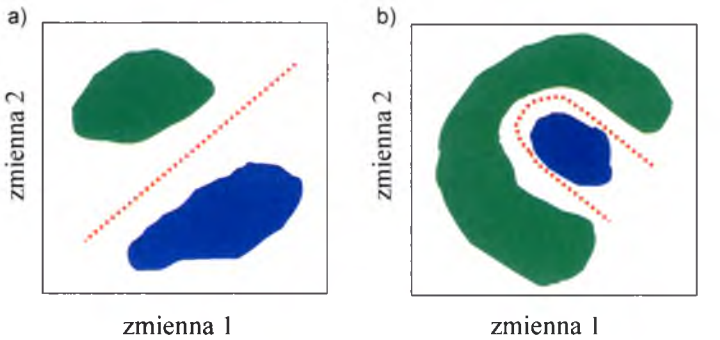
3.3 Wady i zalety metody CART

Modelując dane można spotkać się z różnym rozkładem grup w przestrzeni pomiarowej. Wraz ze wzrostem skomplikowania tego rozmieszczenia rośnie liczba podziałów przestrzeni danych tworzonych w modelu CART. Oznacza to także większą złożoność reguł logicznych modelu CART. Na Rys. 4, w postaci przerywanych linii, zaznaczono podziały przestrzeni danych tworzone przez model CART, w zależności od rozmieszczenia grup obiektów w przestrzeni pomiarowej.



Rys. 4 Podział przestrzeni danych w metodzie CART w zależności od ułożenia grup w tejże przestrzeni

Metody takie jak PLS czy ANN umożliwiają bardziej efektywny podział przestrzeni danych (Rys. 5), jednakże nie pozwalają na konstrukcję reguł logicznych.



Rys. 5 Podział przestrzeni danych dokonany z zastosowaniem metody: a) PLS, b) ANN

4 Cel pracy

Drzewa klasyfikacji i regresji są stosunkowo prostą metodą modelowania danych pozwalającą na konstrukcję reguł logicznych, ale często nie prowadzą do modelu o optymalnej mocy predykcyjnej. Zwykle, powszechnie stosowane w chemii metody modelowania danych takie jak metoda PLS czy sztuczne sieci neuronowe (ANN) [5, 6], pozwalają na konstrukcję modeli obarczonych mniejszym błędem niż metoda CART. Jednakże z drugiej strony sieci neuronowe, które pozwalają na osiągnięcie bardzo dobrych wyników modelowania, dostarczają model którego interpretacja nie jest już tak łatwa jak w przypadku modelu CART. Dlatego pożądana byłaby alternatywna metoda pozwalająca na budowę efektywnego modelu, która dawałaby jednocześnie reguły logiczne. Taką metodą wydają się być neuronowe układy rozmyte, w skrócie NFS [7, 8, 9].

Celem mojej pracy było:

– Zapoznanie się z metodą neuronowych systemów rozmytych oraz jej dostępnymi algorytmami obliczeniowymi.

– Przegląd literaturowy dotychczasowych zastosowań NFS w chemii oraz innych dziedzinach nauki i techniki.

– Porównanie efektywności metody NFS z efektywnością powszechnie stosowanych metod modelowania takich jak PLS, ANN oraz CART dla danych chemicznych o różnej strukturze i wymiarowości.

– Porównanie efektywności metody NFS dla skompresowanych danych oraz tak zwanych zmiennych istotnych.

– Ocena możliwości zastosowania NFS w chemii i korzyści z tego płynących.

5 Systemy wnioskowania rozmytego

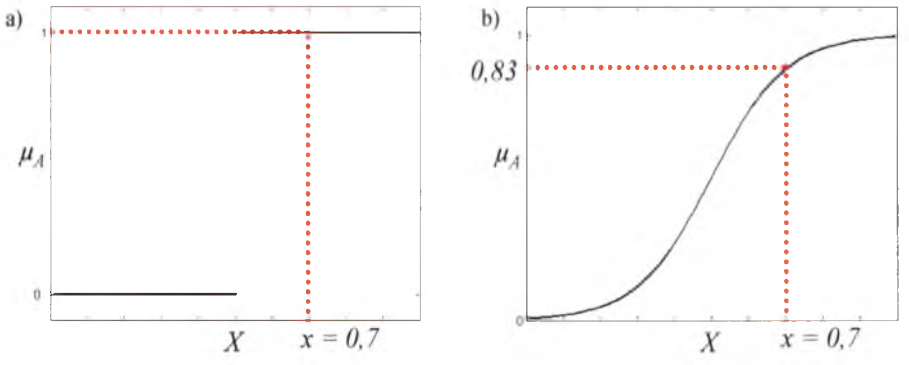
Działanie systemów wnioskowania rozmytego (z ang. *fuzzy inferencje systems*, FIS) [7, 8] opiera się na teorii zbiorów rozmytych oraz ściśle z nią powiązanej logice rozmytej zaprezentowanych przez Zadeha w latach sześćdziesiątych XX wieku [10]. Logika rozmyta to wielowartościowa logika o nieskończonej liczbie stopni prawdziwości stwierdzeń. Jest to obszerny system aksjomatyczny umożliwiający jakościowy oraz ilościowy opis problemów obarczonych różnorodnymi błędami, charakteryzujących się brakiem precyzji czy jednoznaczności. Teoria zbiorów rozmytych dostarcza narzędzi matematycznych pozwalających opisać i analizować takie dane.

5.1 Teoria zbiorów rozmytych

Teoria zbiorów rozmytych pozwala na formalny opis wieloznacznych oraz nieprecyzyjnych zjawisk. W klasycznej teorii zbiorów, która jest dwuwartościowa, obiekt może należeć lub nie do danego zbioru obiektów. Natomiast w teorii zbiorów rozmytych ten sam obiekt może należeć do zbioru, może do niego nie należeć, lub należeć tylko w pewnym stopniu.

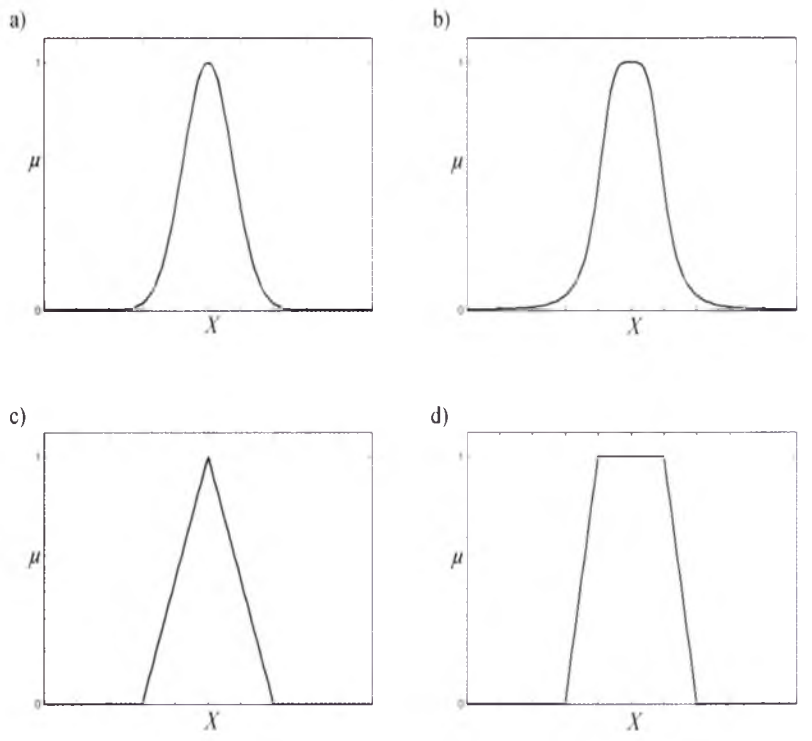
Z matematycznego punktu widzenia przynależność obiektu do danego zbioru opisuje tzw. funkcja przynależności. Funkcje przynależności wykorzystywane są zarówno w klasycznej teorii zbiorów jak i w teorii zbiorów rozmytych.

W klasycznej teorii zbiorów funkcja przynależności obiektu do zbioru jest nieciągła (Rys. 6a) i może przyjmować tylko dwie wartości 0 i 1. „0" oznacza, iż obiekt jest poza zbiorem, a „1" oznacza, że należy do danego zbioru. Taka funkcja przynależności nazywana jest binarną lub singletonem. W teorii zbiorów rozmytych funkcja przynależności obiektu do danego zbioru jest funkcją ciągłą (np. Rys. 6b), przyjmującą wartości od 0 do 1. Jeżeli np. wartość mierzonego parametru $x = 0{,}7$ to stopień przynależności obiektu do zbioru A określony przez funkcję przynależności μ_A wynosi 0,83 i jest jednoznaczny ze stopniem prawdziwości zdania: *Obiekt x należy do zbioru A*; gdzie $x \in X$.



Rys. 6 a) Binarna funkcja przynależności stosowana w klasycznej teorii zbiorów, b) sigmoidalna funkcja przynależności stosowana w teorii zbiorów rozmytych; przerywana linia przedstawia graficzny sposób wyznaczania wartości odpowiedzi funkcji przynależności, μ, dla wartości parametru $x = 0,7$

W teorii zbiorów rozmytych funkcje przynależności mogą przyjmować różne kształty. Na Rys. 6b pokazana jest jedna z najbardziej popularnych funkcji przynależności tzw. sigmoidalna funkcja przynależności. Inne powszechnie stosowane funkcje przynależności przedstawia Rys. 7. Decyzja o wyborze kształtu funkcji przynależności jest arbitralna i zależy od rodzaju problemu, będącego przedmiotem badań.



Rys. 7 Przykładowe funkcje przynależności stosowane w teorii zbiorów rozmytych: a) funkcja Gaussa, b) funkcja dzwonowa, c) funkcja trójkątna oraz d) funkcja trapezoidalna

Funkcje te opisane są następującymi wzorami [8]:

– binarna funkcja przynależności

$$\mu_{\text{Sing}}(x) = \begin{cases} 1, x \in A \\ 0, x \notin A \end{cases}; \tag{4}$$

gdzie: x to wartość mierzonego parametru, A oznacza zbiór;

– sigmoidalna funkcja przynależności

$$\mu_{\text{Sig}}(x) = \frac{1}{1 + \exp[-a(x-c)]}, \tag{5}$$

gdzie a to nachylenie funkcji dla punktu przegięcia c;

– funkcja przynależności Gaussa

$$\mu_{\text{Gaus}}(x) = \exp\left(-\left(\frac{x-\bar{x}}{\sigma}\right)^2\right), \tag{6}$$

gdzie: $\bar{x}$ to wartość położenia maksimum piku Gaussa natomiast σ jest parametrem określającym szerokość krzywej Gaussa;

– dzwonowa funkcja przynależności

$$\mu_{\text{Bell}}(x) = \frac{1}{1 + \left|\frac{x-c}{a}\right|^{2b}}, \tag{7}$$

gdzie: c to położenie środka funkcji μ_{Bell}, a i b to parametry określające odpowiednio szerokość funkcji oraz jej punkty przegięcia;

– trójkątna funkcja przynależności

$$\mu_{\text{Tri}}(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{b-a}, & a \le x \le b \\ \frac{c-x}{x-b}, & a \le x \le c \\ 0, & c \le x \end{cases} \tag{8}$$

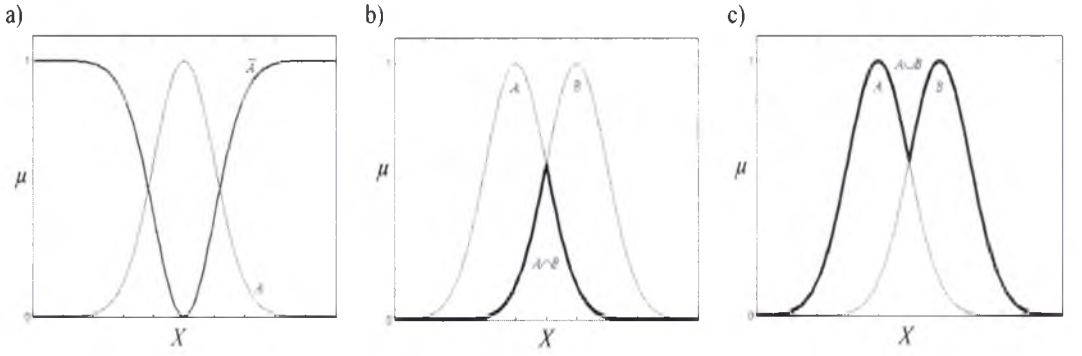
gdzie: b definiuje położenie maksimum funkcji trójkątnej natomiast a oraz c to parametry określające szerokość funkcji μ_{Tri};

– trapezoidalna funkcja przynależności

$$\mu_{\text{Trap}}(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \dfrac{d-x}{d-c}, & c \le x \le d \\ 0, & d \le x \end{cases} \qquad 9$$

gdzie: $a<b<c<d$ to parametry opisujące kolejno zmiany monotoniczności funkcji μ_{Trap}.

Operacje matematyczne na zbiorach rozmytych wykonuje się na funkcjach przynależności [8]. Podstawowe operacje to dopełnienie zbioru, suma zbiorów, przecięcie zbiorów oraz iloczyn algebraiczny. Graficzna ilustracja tych operacji przedstawiona jest na Rys. 8.



Rys. 8 Podstawowe operacje matematyczne wykonywane na zbiorach rozmytych z wykorzystaniem funkcji przynależności: a) dopełnienie zbioru, b) przecięcie dwóch zbiorów, c) suma dwóch zbiorów; gdzie linią pogrubioną zaznaczony jest rezultat każdej z wykonanych operacji

– Dopełnienie zbioru A, $\overline{A}$ (z ang. *complement*), które oblicza się jako jeden minus wartość funkcji przynależności (Rys. 8a) przedstawia następujące równanie:

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x), \ \forall x \in X; \qquad 10$$

– Przecięcie zbiorów A oraz B, $A \cap B$ (z ang. *intersection*, Rys. 8b), jest równe minimum z obu funkcji przynależności:

$$\mu_{A \cap B}(x) = min[\mu_A(x), \mu_B(x)] \equiv \mu_A(x) \wedge \mu_B(x), \ \forall x \in X; \qquad 11$$

gdzie: $A \cap B \subseteq A$ oraz $A \cap B \subseteq B$.

– Iloczyn algebraiczny zbiorów A oraz B, $A \bullet B$ (z ang. *algebraic produkt*), jest równy:

$$\mu_{A \bullet B}(x) = \mu_A(x) \bullet \mu_B(x), \ \forall x \in X; \tag{12}$$

– Suma dwóch zbiorów, $A \cup B$ (z ang. *union*, Rys. 8c), to maksimum dwóch funkcji przynależności:

$$\mu_{A \cup B}(x) = max[\mu_A(x), \mu_B(x)] \equiv \mu_A(x) \vee \mu_B(x), \ \forall x \in X; \tag{13}$$

gdzie: $A \subseteq A \cup B$ oraz $B \subseteq A \cup B$.

Inne warte odnotowania, choć mniej popularne, matematyczne definicje wykorzystywane w teorii zbiorów rozmytych to: równość zbiorów, stopień równości zbiorów, wydzielenie podzbioru z innego zbioru, stopień zawierania się podzbioru w zbiorze oraz podwójna negacja.

– Równość zbioru A oraz B, $A = B$ (z ang. *equality*), oznacza, iż zbiory A oraz B zawierają te same obiekty:

$$\mu_A(x) = \mu_B(x), \ \forall x \in X; \tag{14}$$

– Stopień równości zbiorów A oraz B, $E(A,B)$ (z ang. *overlapping degree*), określa stopień w jakim zbiór A nachodzi na zbiór B:

$$E(A,B) = \frac{|A \cap B|}{|A \cup B|}; \tag{15}$$

gdzie $0 \le E(A,B) \le 1$.

– zbiór A jest podzbiorem zbioru B, $A \subseteq B$ (z ang. *subset*), wtedy i tylko wtedy gdy wszystkie obiekty ze zbioru A należą do zbioru B:

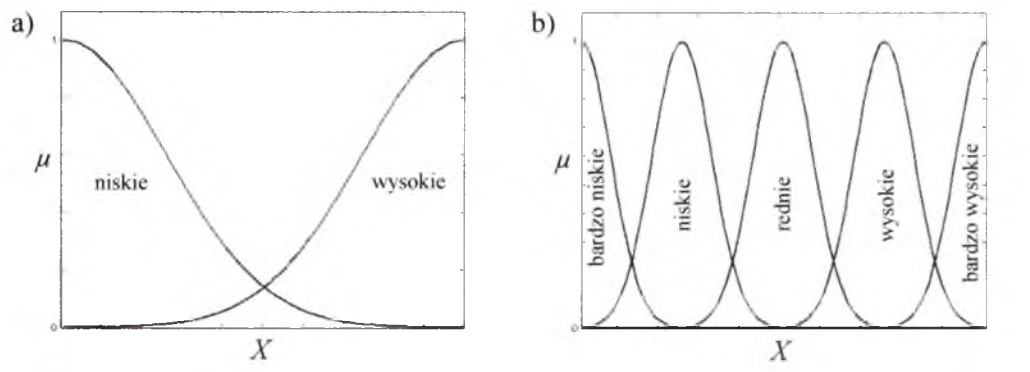
$$\mu_A(x) \le \mu_B(x), \ \forall x \in X; \tag{16}$$

– Stopień podzbiorowości, $S(A, B)$ (z ang. *subsethood measure*), określa stopień w jakim zbiór A jest podzbiorem zbioru B:

$$S(A,B) \equiv \text{degree}(A \subseteq B) = \frac{|A \cap B|}{|A|}. \tag{17}$$

Powyżej zamieszczono tylko podstawowe definicje wykonywane w teorii zbiorów rozmytych, więcej szczegółów można znaleźć w [8].

Przestrzeń pomiarowa obejmująca zakres zmienności mierzonych parametrów, może być opisana przez różną liczbę funkcji przynależności. Jest to równoznaczne z podziałem przestrzeni pomiarowej na daną liczbę podzbiorów rozmytych [7].



Rys. 9 Podział przestrzeni pomiarowej X przez wprowadzenie a) dwóch oraz b) pięciu funkcji przynależności, odpowiadających różnym zakresom wartości mierzonego parametru X

Rys. 9 przedstawia przykładowy podział przestrzeni pomiarowej z zastosowaniem odpowiednio dwóch i pięciu funkcji przynależności. Zastosowanie dwóch funkcji przynależności pozwala na utworzenie dwóch rozmytych zbiorów w przestrzeni pomiarowej, jednego odpowiadającego niskim wartościom mierzonego parametru oraz drugiego odpowiadającego jego wysokim wartościom (Rys. 9a). Natomiast Rys. 9b przedstawia podział przestrzeni pomiarowej na pięć zbiorów rozmytych zawierających obiekty odpowiednio o bardzo niskich, niskich, średnich, wysokich oraz bardzo wysokich wartościach mierzonego parametru X. Ilość zastosowanych funkcji przynależności jest, podobnie jak ich kształt, decyzją arbitralną i dostosowaną do konkretnego problemu badawczego.

5.2 Reguły logiczne oraz wnioskowanie rozmyte

Reguły logiczne ułatwiają interpretację modeli matematycznych. Dlatego pożądane są techniki modelowania danych pozwalające na konstrukcję reguł logicznych jak np. model systemów wnioskowania rozmytego.

W klasycznej teorii zbiorów wnioskuje się o prawdzie bądź fałszu następnika B, w oparciu o wartość logiczną poprzednika, A. Wnioskowanie to odbywa się zgodnie z regułą *modus ponens*: $A \rightarrow B$ $((A \wedge (A \rightarrow B)) \rightarrow B)$ [9, 11]. Zgodnie z tą regułą, jeśli $A = 1$ to także i $B = 1$. *Modus ponens* przedstawia poniższy schemat:

Reguła:	JEŻELI *x należy do A* TO *y należy do B*
Obserwacja:	*x należy do A*
Wniosek:	*y należy do B*

Najprostszym przykładem reguły logicznej może być zdanie:

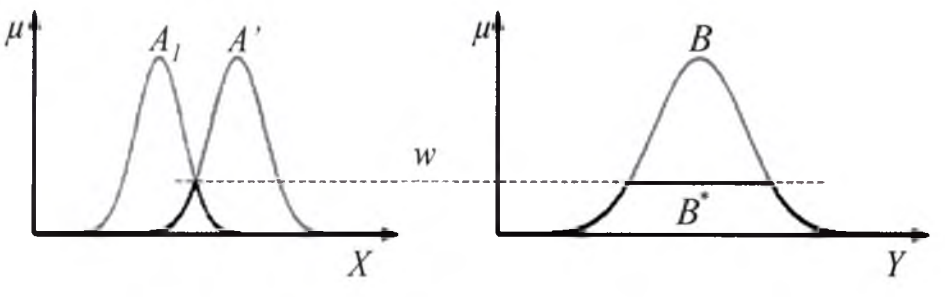
– JEŻELI *papierek wskaźnikowy jest niebieski* TO *roztwór jest zasadowy*.

Jest to reguła logiczna typu JEŻELI-TO, którą w formie skróconej zapisuje się następująco: $A \rightarrow B$. Reguła taka składa się odpowiednio z części poprzednika A oraz następnika B oddzielonych zaimkiem TO.

Wnioskowanie rozmyte odbywa się oparciu o rozszerzony na przypadek rozmyty *modus ponens*, tzw. uogólniony *modus ponens* (z ang. *generalized modus ponens*, GMP). Postrzeganie kolorów przez człowieka jest sprawą subiektywną, np. niebieski papierek może być jaśniejszy lub ciemniejszy, a roztwór może być bardziej lub mniej zasadowy. Dlatego w teorii zbiorów rozmytych prawdziwość poprzednika oraz następnika określana jest nie w systemie prawda/fałsz (0 lub 1), ale w systemie liczb ciągłych od 0 do 1 przez zastosowanie odpowiedniej ciągłej funkcji przynależności. Funkcja przynależności określa stopień prawdziwości poprzednika i decyduje o tzw. stopniu odpalenia reguły logicznej, czyli stopniu prawdziwości następnika. Schemat wnioskowania rozmytego przedstawić można więc następująco (Rys. 10):

Reguła:	JEŻELI x należy do A TO y należy do B
Obserwacja:	x należy do A'
Wniosek:	y należy do B'

gdzie: A' oraz B' oznaczają zbiory rozmyte bliskie/podobne zbiorom rozmytym A oraz B, będące odpowiednio tzw. wejściem rozmytym oraz wyjściem rozmytym. Każdy z tych zbiorów opisany jest odpowiednią funkcją przynależności, $\mu_{A'}$ oraz $\mu_{B'}$. Wartość prawdy poprzednika reguły logicznej, w (Rów. 18), dla GMP prezentowanego na Rys. 10 definiowana jest jako maksimum podzbioru utworzonego przez przecięcie zbioru A oraz A'. Maksimum to jednocześnie determinuje stopień prawdziwości następnika, a więc i reguły logicznej.



Rys. 10 Schemat wnioskowania z zastosowaniem reguły uogólnionego *modus ponens*, GMP

$$w = \max[\mu_A \wedge \mu_{A'}] = \max[\min(\mu_A, \mu_{A'})] \qquad 18$$

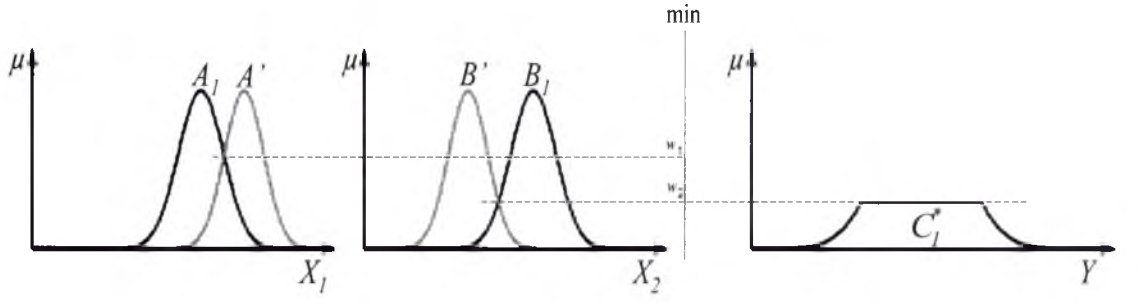
Przykładowe rozmyte reguły logiczne typu JEŻELI-TO mogą mieć następującą postać:

- JEŻELI *wartość temperatury jest wysoka* TO *szybkość reakcji jest duża*;
- JEŻELI *ciśnienie jest duże* TO *objętość jest mała.*

Część poprzednika może być rozbudowana, dlatego istnieje także możliwość wnioskowania w oparciu o więcej niż jedno zdanie logiczne, np.:

- JEŻELI *wartość temperatury jest duża* ORAZ *wartość pH jest mała* LUB *roztwór jest mętny* TO *szybkość reakcji jest duża.*

Spójniki ORAZ i LUB oznaczają operacje logiczne: odpowiednio sumowanie i przecięcie zbiorów. Wartość prawdy poprzednika jest obliczana jako minimum (Rów. 19) lub jako iloczyn (Rów. 20) z wartości prawd poszczególnych zdań składowych (tworzących rozbudowany poprzednik reguły logicznej), w_i,... w_j, gdzie i oraz j odpowiadają kolejnym zdaniom składowym (Rys. 11).

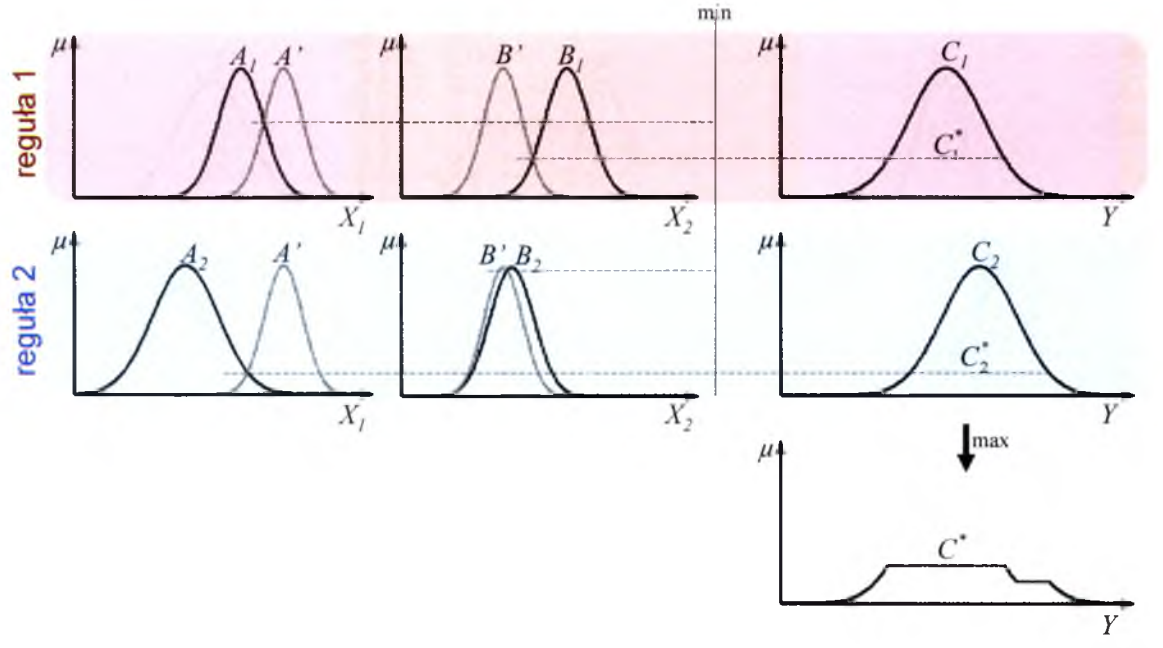


Rys. 11 Schemat GMP na podstawie rozbudowanego poprzednika (dwa zdania składowe) z wykorzystaniem operatora minimum (min) z w_1 i w_2

$$w = \min[w_1, w_2] \tag{19}$$

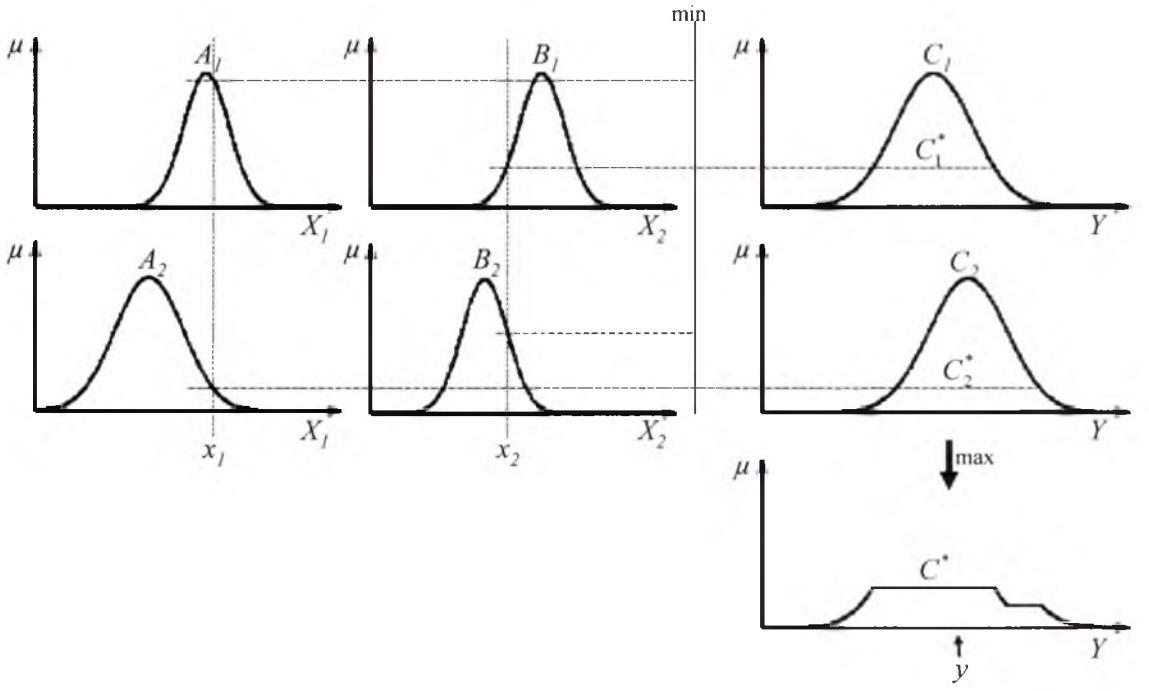
$$w = w_1 \bullet w_2 \tag{20}$$

W ramach wnioskowania rozmytego rozbudowaniu mogą ulec nie tylko same reguły logiczne, ale także ich ilość może zostać zwielokrotniona. Na wyjściu z systemu przedstawionego na Rys. 12 otrzymuje się zbiór rozmyty C^*, obliczany jako maksimum ze zbiorów rozmytych C_1^* i C_2^*, określających stopnie prawdziwości poszczególnych reguł logicznych.



Rys. 12 Schemat wnioskowania rozmytego dla wejścia oraz wyjścia będącego zbiorami rozmytymi A' i B' w przestrzeni pomiarowej odpowiednio X_i oraz Y, z wykorzystaniem operatora minimum (min), gdzie μ oznacza wartości funkcji przynależności: A_i, B_i, C_i, natomiast * oznacza wynik wnioskowania zarówno w obrębie danej reguły jak i całego GMP

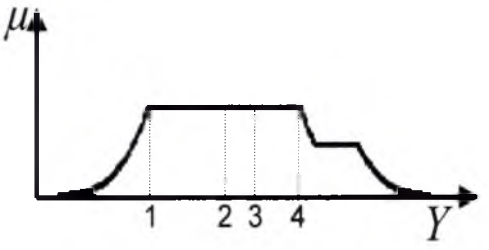
Wnioskowanie rozmyte może być przeprowadzane nie tylko w oparciu o wejścia będące zbiorami rozmytymi, A' czy B', ale także w oparciu o konkretne wartości liczbowe – skalary, x_i (Rys. 13). Skalary x_1 oraz x_2 są wartościami mierzonych parametrów X_1 i X_2. W celu oszacowania wartości prawdy każdego ze składowych poprzedników określa się stopień przynależności x_1 oraz x_2 odpowiednio do podzbiorów A_1, A_2 oraz B_1, B_2. Niech stopień przynależności x_1 do zbioru rozmytego A_1 określony przez odpowiednią funkcję przynależności wynosi 0,95; a do zbioru rozmytego A_2 0,30. Ponadto niech wartości przynależności dla x_2 do zbiorów rozmytych B_1 i B_2 wynoszą odpowiednio 0,20 i 0,60. Stopień prawdziwości poprzednika pierwszej reguły logicznej wyznacza zastosowany operator, dla operatora minimum (min) wartość ta wynosi 0,30 (min[0,30; 0.95]). Podobnie stopień odpalenia drugiej reguły logicznej wyznaczony operatorem minimum wynosi 0,20 (min[0,20; 0,60]). Ostateczne wyjście z GMP, czyli wniosek, ma postać zbioru rozmytego C^*.



Rys. 13 Schemat wnioskowania rozmytego dla wejść będących skalarami x_1 i x_2 oraz rozmytego wyjścia C_i w przestrzeni pomiarowej odpowiednio X_i oraz Y, z wykorzystaniem operatora minimum (min), gdzie μ oznacza wartości funkcji przynależności: A_i, B_i, C_i, a * oznacza wynik wnioskowania zarówno w obrębie danej reguły jak i całego GMP; ponadto y oznacza wyjście będące rezultatem zastosowania procedury wyostrzania

Jeżeli pożądane jest otrzymanie konkretnej wartości liczbowej (z ang. *crisp value*) wymagane jest zastosowanie tzw. procedury wyostrzania – defuzyfikacji. Defuzyfikacja (z ang. *defuzzification*) polega na ekstrahowaniu ze zbioru rozmytego wartości liczbowej – skalaru – w oparciu o dostępne algorytmy. Istnieje wiele algorytmów wyostrzania, a w zależności od zastosowanej metody otrzymane wyniki mogą się różnić w mniejszym lub większym stopniu [7]. Wyboru metody wyostrzania dokonuje się pod kątem analizowanego problemu. Do najpowszechniej stosowanym algorytmów wyostrzania należą (Rys. 14):

- najmniejsza wartość z maksimum (z ang. *smallest of max.*, SOM; lub *first of maximum*, FOM)
- średnia wartość z maksimum (*mean of max.*, MOM)
- centroid powierzchni pod krzywą (z ang. *centre of area*, COA; lub *centra of gravity*, COG)
- największa wartość z maksimum (z ang. *largest of max.*, lub *last of max.*, LOM)



Rys. 14 Przykładowa ilustracja zależności otrzymanego wyniku procedury wyostrzania od zastosowanego algorytmu, gdzie: 1) najmniejsza wartość z maksimum, 2) średnia wartość z maksimum, 3) centroid powierzchni pod krzywą, 4) największa wartość z maksimum

Istnieją także inne, mniej znane algorytmy wyostrzania, których opis można znaleźć w [12].

5.3 Typy systemów wnioskowania rozmytego

Zbiory rozmyte, jak i operacje na nich wykonywane oraz reguły logiczne typu JEŻELI-TO, stanowią rdzeń systemów wnioskowania rozmytego (FIS), których atrakcyjność polega na możliwości modelowania małoprecyzyjnych danych. FIS to multidyscyplinarna technika przetwarzania danych i z tego też powodu w literaturze opisywana jest pod wieloma nazwami, np.: *fuzzy-rule-based system*, *fuzzy expert system*, *fuzzy associative memory* czy *fuzzy system*.

Opracowano wiele systemów wnioskowania rozmytego, wśród których pomimo różnorodności można wyróżnić następujące elementy składowe:

- baza rozmytych reguł logicznych typu JEŻELI-TO;
- baza funkcji przynależności, zdefiniowanych i wykorzystywanych w bloku reguł;
- mechanizm wnioskowania umożliwiający przeprowadzenie procedury wnioskowania w oparciu o reguły i dostępne dane.

FIS akceptują na wejściu zarówno informację w postaci zbiorów rozmytych jak i skalarów. W większości przypadków jednak wyjście z systemu wnioskowania rozmytego ma postać rozmytą, a uzyskanie skalaru wymaga zastosowania procedury wyostrzania. W ramach systemów wnioskowania rozmytego sygnał przekazywany jest od wejścia do wyjścia, przy czym dostępna na wyjściu informacja jest nieliniową konsekwencją informacji podanej na wejście. Można mówić o swoistym mapowaniu przestrzeni danych przez FIS, które następuje w oparciu o rozmyte reguły typu JEŻELI-TO. W regule logicznej każdy z poprzedników opisuje, poprzez odpowiednią funkcję przynależności, pewien konkretny obszar wejściowej przestrzeni pomiarowej. Podobnie rzecz ma się w przypadku następników, z tą tylko różnicą, iż odnoszą się one do przestrzeni wyjściowej.

Poniżej przedstawiono trzy najpopularniejsze systemy wnioskowania rozmytego, które znalazły najwięcej zastosowań w różnych dziedzinach nauki i techniki.

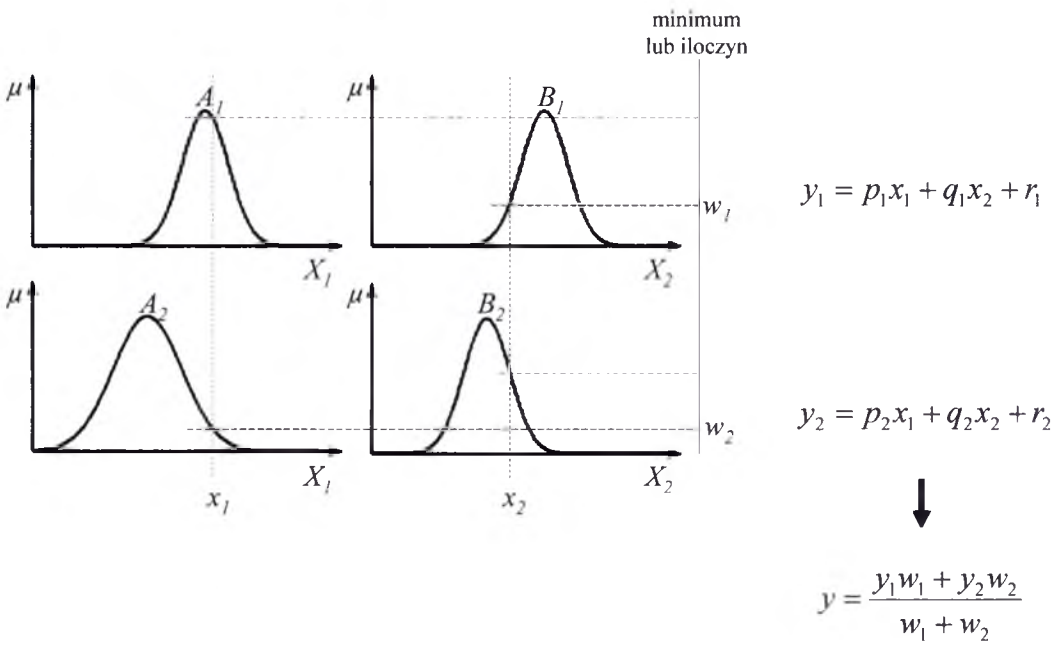
Prezentowane FIS różnią się w swojej budowie w części następnika rozmytych reguł logicznych. Skutkiem występujących różnic w budowie poszczególnych systemów wnioskowania rozmytego jest inny sposób agregacji i wyostrzania sygnału w ramach systemów.

5.3.1 System wnioskowania rozmytego typu Mamdani

System wnioskowania rozmytego typu Mamdani został opracowany w latach siedemdziesiątych XX wieku do kontroli układu silnika parowego i bojlera [13, 14, 15]. Na Rys. 13 widoczne jest, iż na wyjściu system dostarcza rozmytej odpowiedzi w postaci zbioru rozmytego, C^*. Celem uzyskania jednoznacznej wartości y, konieczne jest zastosowanie procedury wyostrzenia, której dobór zależny jest od analizowanego problemu.

5.3.2 System wnioskowania rozmytego typu Takagi, Sugeno i Kang

Drugim omawianym systemem wnioskowana jest system opracowany przez T. Takagi, M. Sugeno oraz G.T. Kanga w latach osiemdziesiątych ubiegłego wieku [16, 17] i znany jako Sugeno FIS.



Rys. 15 System wnioskowania rozmytego typu Sugeno pierwszego rzędu dla wejścia w przestrzeni pomiarowej X_i oraz wyjścia będących skalarami, gdzie μ oznacza wartości danych funkcji przynależności: A_i, B_i; y_i to liniowe kombinacje mierzonych parametrów, natomiast y oznacza wyjście będące ważoną sumą odpowiadających sobie stopni odpalenia poprzedników reguł logicznych w_i oraz liniowych kombinacji oryginalnych zmiennych

W ramach systemu wnioskowania rozmytego typu Sugeno następnik każdej reguły logicznej ma formę liniowej kombinacji oryginalnych zmiennych X_1 i X_2. Ostateczną odpowiedzią systemu wnioskowania rozmytego typu Sugeno (y) jest suma linowych kombinacji oryginalnych zmiennych ważona stopniami prawdziwości odpowiadających im poprzedników reguł logicznych (w_1 i w_2) obliczana według wzoru z rysunku 15.

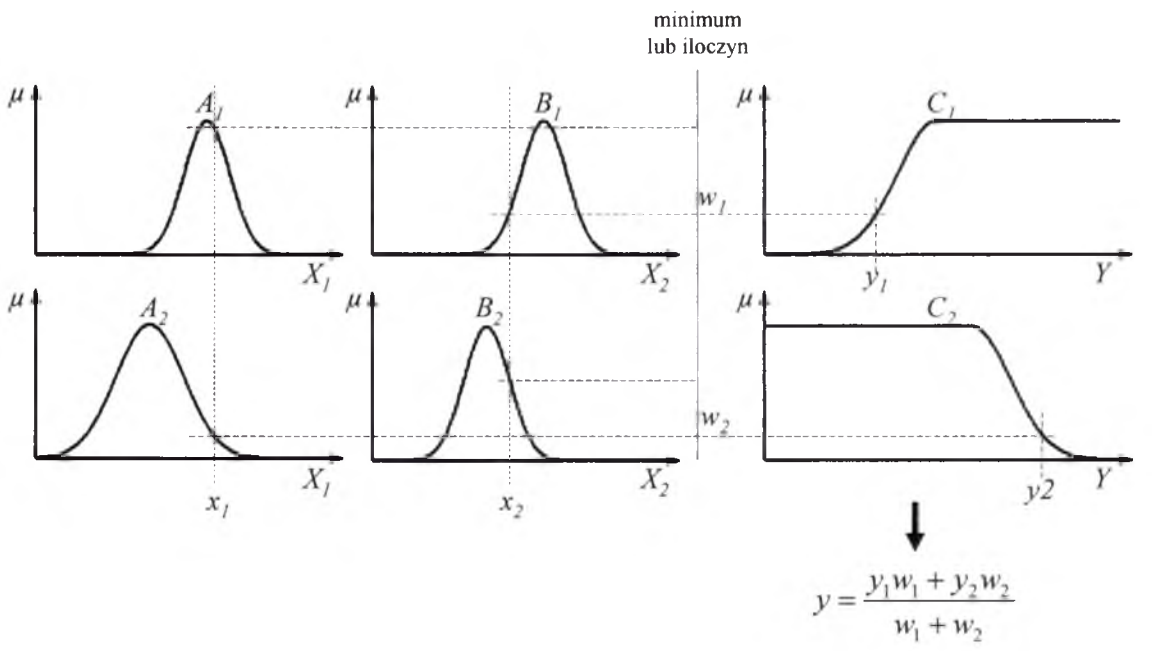
Typową rozmytą regułę logiczną typu JEŻELI-TO konstruowaną w ramach Sugeno FIS można zapisać następująco:

- JEŻELI x_1 należy do A ORAZ x_2 należy do B TO $y = f(x_1, x_2)$,

gdzie: A oraz B to zbiory rozmyte w ramach poprzednika. Funkcja $f(x_1, x_2)$ stanowi następnik i z reguły jest wielomianem zerowego lub pierwszego rzędu. Mówi się wtedy odpowiednio o systemie wnioskowania rozmytego typu Sugeno zerowego rzędu lub pierwszego rzędu. Decyzja o rzędowości stosowanego Sugeno FIS jest zależna od analizowanego problemu. Ponadto, system wnioskowana rozmytego typu Sugeno zerowego rzędu może być traktowany jako specjalny przypadek systemu wnioskowania rozmytego typu Mamdani. Warunkiem ekwiwalencji obu systemów jest zastosowanie funkcji przynależności typu singleton (Rys. 6a) w ramach następnika systemu Mamdani FIS. Główną zaletą systemu wnioskowania typu Sugeno w porównaniu z systemem typu Mamdani jest ostre wyjście.

5.3.3 System wnioskowania rozmytego typu Tsukamoto

Trzecim omawianym systemem wnioskowani rozmytego jest Tsukamoto FIS [18]. System ten został opracowany pod koniec lat siedemdziesiątych XX wieku przez Y. Tsukamoto. Następniki reguł logicznych w Tsukamoto FIS stanowią monotoniczne funkcje przynależności pozwalające uzyskać wyjście z danej reguły logicznej w formie ostrej liczb y_i (skalaru). Ostateczna odpowiedź systemu wnioskowania rozmytego typu Tsukamoto jest obliczana w sposób analogiczny do systemu wnioskowania rozmytego typu Sugeno (Rys. 15). Przedstawiony na rysunku 16 Tsukamoto FIS ma wejście w postaci skalarnej. Istnieje jednak możliwość, aby wejście miało formę zbioru rozmytego (analogicznie do Rys. 12).



Rys. 16 Schemat systemu wnioskowania rozmytego typu Tsukomoto dla wejścia oraz wyjścia będącego skalarami w przestrzeni pomiarowej odpowiednio X_i oraz Y, gdzie μ oznacza wartości danych funkcji przynależności: A_i, B_i; C_i; y oznacza wyjście będące ważoną sumą odpowiadających sobie stopni odpalenia poprzedników reguł logicznych (w_i) oraz stopni prawdziwości reguł logicznych (y_i)

5.4 Zastosowania systemów wnioskowania rozmytego

Logika rozmyta oraz systemy wnioskowania rozmytego (FIS) znajdują zastosowanie w różnorodnych gałęziach nauki oraz przemysłu. FIS są integralną częścią wielu systemów eksperckich (np. [19]), a także systemów kontroli procesów (np. [20]). Systemy wnioskowania rozmytego znalazły także zastosowanie w wielu urządzeniach RTV i AGD [21,22], w medycynie [23,24]; w kryminalistyce do przetwarzania obrazów [25,26] oraz w prognozowaniu pogody [27].

Zalety wnioskowania rozmytego zostały dostrzeżone także w chemii oraz inżynierii chemicznej, głównie w kontroli procesów i reakcji chemicznych. Aplikacje FIS można znaleźć również w przemyśle farmaceutycznym. Przykładem takiego zastosowania może być kontrola reakcji chemicznych na skalę przemysłową [28,29], która bardzo często prowadzona jest w czasie rzeczywistym, w tzw. systemie *on-line* [30]. Więcej przykładów przemysłowych zastosowań można znaleźć w następujących publikacjach [31,32,33,34,35].

Logika rozmyta i wnioskowanie rozmyte umożliwia porównanie i obróbkę różnego rodzaju sygnałów i danych chemicznych, takich jak widma w bliskiej podczerwieni [36,37,38,39], chromatogramy [40,41], czy elektroforegramy [42]. Teoria zbiorów rozmytych w ramach FIS znalazła także zastosowanie jako metoda klasyfikacji oraz kalibracji [43,44,45,46,47].

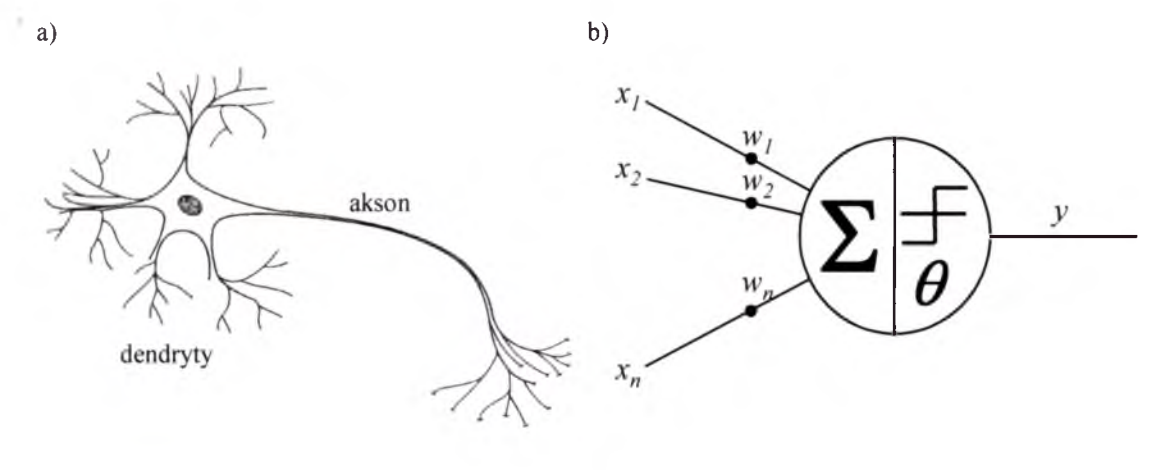
5.5 Wady i zalety systemów wnioskowania rozmytego

Systemy wnioskowania rozmytego umożliwiają modelowanie nieprecyzyjnych danych i kontrolę różnorodnych procesów oraz łatwą interpretację modelu. FIS należą do tzw. grupy systemów eksperckich i jako takie zostały zaprojektowane tak by można było wykorzystać wiedzę eksperta do konstrukcji modelu. Jednak nie zawsze posiadana wiedza pozwala na określenie optymalnej struktury modelu, czyli na zdefiniowanie ilości i kształtu funkcji przynależności. Potrzeba więc strategii, która pozwoliłaby na automatyczny dobór parametrów konstruowanego modelu w oparciu o analizowane dane. Tu z pomocą przychodzą sieci neuronowe, które mają zdolność adaptacji do struktury danych.

6 Sieci neuronowe

Pod terminem sieci neuronowe kryją się systemy przetwarzania informacji, których nazwa nawiązuje do sposobu przesyłu i przetwarzania informacji przez ludzki układ nerwowy [6]. Swoje właściwości sieci neuronowe zawdzięczają równoległemu sposobi przetwarzania informacji – to właśnie tam kryje się moc sieci.

Komórka nerwowa, neuron (Rys. 17a), posiada jądro komórkowe przetwarzające informacje dostarczane przez dendryty. Po przetworzeniu informacji w jądrze komórkowym, informacja przekazywana jest do kolejnej komórki lub grupy komórek poprzez akson.



Rys. 17 a) Komórka nerwowa, b) sztuczny neuron autorstwa McCullocha i Pittsa (neuron M-P), gdzie x_i oznacza i-te wejście, w_i stowarzyszoną z nim wagę, y wyjście z neuronu, natomiast w węźle zamieszczone są odpowiednio symbole operacji sumowania oraz funkcji aktywacji neuronu i wartości progowej

W roku 1943 McCulloch i Pitts opracowali matematyczny model sztucznego neuronu (Rys. 17b), tzw. neuron M-P. Model ten w żadnym razie nie jest matematycznym odzwierciedleniem sposobu działania naturalnej komórki nerwowej, a jedynie powstał nią zainspirowany [48]. Informacja dostarczana jest do sztucznego neuronu przez wejścia x_i, a każdemu wejściu przypisana jest waga w_i (Rys. 17b). Wartość wagi mówi o sile wejścia, dodatni znak oznacza wzmocnienie sygnału, a ujemny jego osłabienie. Wejście do neuronu może także zostać wyłączone gdy odpowiedniej wadze przypisana zostanie wartość równa zeru. Ważone wejścia

są sumowane i przetwarzane przez funkcję aktywacji neuronu typu singelton (Rys. 6a, Rów. 4). Wyjście z neuronu M-P, obliczane jest według wzoru:

$$y = \mu\left(\sum_{i=1}^{m} w_i x_i - \theta\right) \qquad 21$$

gdzie; w_i to waga odpowiadająca i-temu wejściu x_i, a θ to wartość progowa aktywacji neuronu.

Przez dobór odpowiednich wag neuron M-P umożliwia wykonanie prostych działań klasyfikacyjnych.

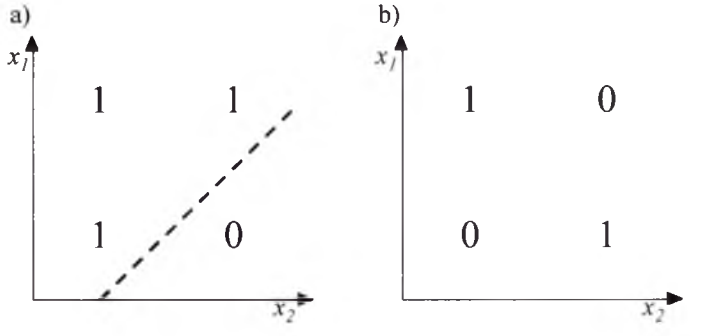
Zestawienie kilku neuronów pozwala otrzymać sieć. Pierwszą opublikowaną siecią neuronową jest tzw. perceptron autorstwa F. Rosenblatta [49,50]. Jednak jest to sieć, która umożliwia rozwiązanie tylko liniowych problemów dyskryminacyjnych. Perceptron nie umożliwia więc rozwiązania tzw. problemu XOR (Rys. 18b) [51]. Wskazują na to w swej książce M. Minsky i S. Papert dowodząc, iż jednowarstwowe sieci wyposażone w nieciągłą funkcję aktywacji (np.: Rów. 22 i 23) mają bardzo ograniczone zastosowania. Przekonanie o ograniczeniach w zastosowaniach perceptronu obalili J.A. Andersen oraz E. Rosenfeld w roku 1988 wprowadzając ciągłą funkcję aktywacji [52]. Poniżej zamieszczono oryginalne funkcje aktywacji wykorzystane w perceptronie:

- funkcja skoku jednostkowego bipolarna

$$f(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}; \qquad 22$$

- funkcja skoku jednostkowego unipolarna

$$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}. \qquad 23$$



Rys. 18 Przykładowe rozmieszczenie obiektów w dwuwymiarowej przestrzeni danych: a) problem separowalny liniowo; b) problem nieseparowalny linowo, tzw. problem XOR

Kolejną siecią neuronową, o której warto wspomnieć była pierwsza oferowana komercyjnie sieć o nazwie Adeline (z ang. *adaptive linear network*) [6]. Znalazła ona zastosowanie miedzy innymi w telekomunikacji i w przemyśle obronnym w urządzeniach radarowych [48]. Opis innych sieci można znaleźć w [53].

6.1 Rodzaje sieci neuronowych

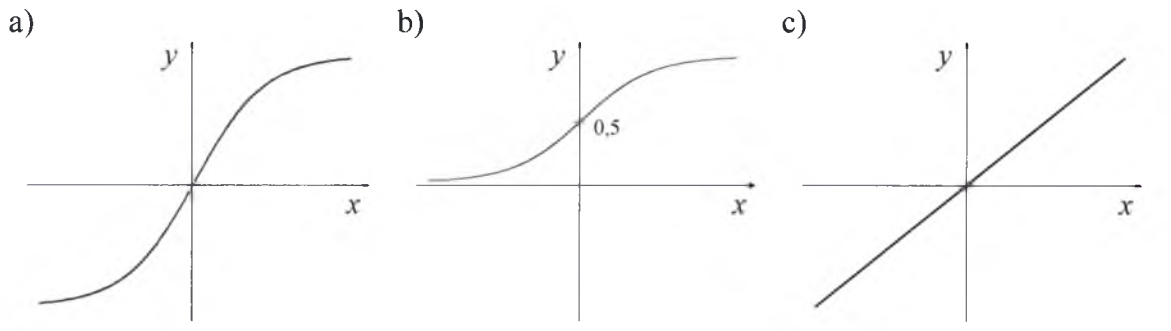
Istnieje wiele kryteriów podziału sieci neuronowych [5]. Ze względu na architekturę, sieci neuronowe można podzielić na jednowarstwowe oraz wielowarstwowe. Stosując jako kryterium sposób przetwarzania informacji przez sieć, można wyróżnić sieci z połączeniami jednokierunkowymi (z ang. *feedforward*) i sieci ze sprzężeniami zwrotnymi (np. sieci Hopfielda). Za pomocą sieci neuronowych można realizować proces uczenia zarówno z nadzorem, zwany także uczeniem z nauczycielem (z ang. *supervised learning*), jak też uczenie bez nadzoru, czyli bez nauczyciela (z ang. *unsupervised learning*).

Ze względu na temat niniejszej pracy zostaną omówione tu tylko jednokierunkowe sieci neuronowe przystosowane do uczenia z nauczycielem (z ang. *supervised feedforward neural networks*)

6.2 Funkcje aktywacji neuronu

Podstawowym elementem budulcowym sieci neuronowej jest sztuczny neuron. Schematycznie budowę takiego neuronu przedstawia Rys. 17b. Funkcja aktywacji sztucznego neuronu (z ang. *activation function* lub *transfer function*) może mieć charakter liniowy lub nieliniowy (Rys. 19). Ponadto funkcja taka musi spełniać następujące warunki:

- ciągłość pomiędzy wartością minimalną a maksymalną funkcji;
- łatwość obliczenia i ciągłość pochodnej funkcji;
- możliwość wprowadzenia do argumentu funkcji parametru modyfikującego kształt krzywej.



Rys. 19 Przykładowe funkcje aktywacji neuronu sieci neuronowej: a) tangens hiperboliczny, b) funkcja sigmoidalna, c) funkcja liniowa

Wybór funkcji aktywacji neuronu jest uzależniony od zadania, jakie będzie on wypełniał (kalibracja lub klasyfikacja). Do najpowszechniej stosowanych funkcji aktywacji neuronu (Rys. 19) należą tangens hiperboliczny, funkcja sigmoidalna oraz funkcja liniowa (odpowiednio Rów. 24-26). Funkcje te opisane są następującymi wzorami:

- tangens hiperboliczny

$$f(x) = \frac{1 - e^{-\beta x}}{1 + e^{-\beta x}}, \ \beta > 0, \tag{24}$$

gdzie:β to współczynnik określający nachylenie funkcji, które rośnie wraz ze wzrostem tego współczynnika;

- funkcja sigmoidalna, nazywana także krzywą logistyczną lub signum

$$f(x) = \frac{1}{1 + e^{-\beta x}}, \ \beta > 0; \tag{25}$$

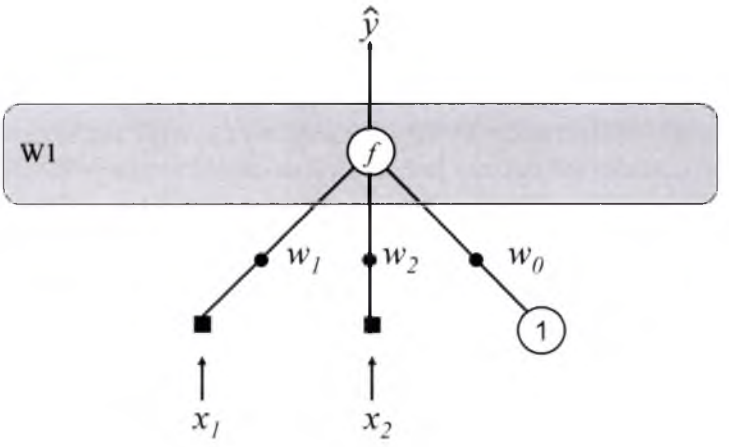
- funkcja liniowa

$$f(x) = ax + b \tag{26}$$

gdzie: a oraz b to parametry funkcji nazywane odpowiednio współczynnikiem kierunkowym funkcji i wyrazem wolnym.

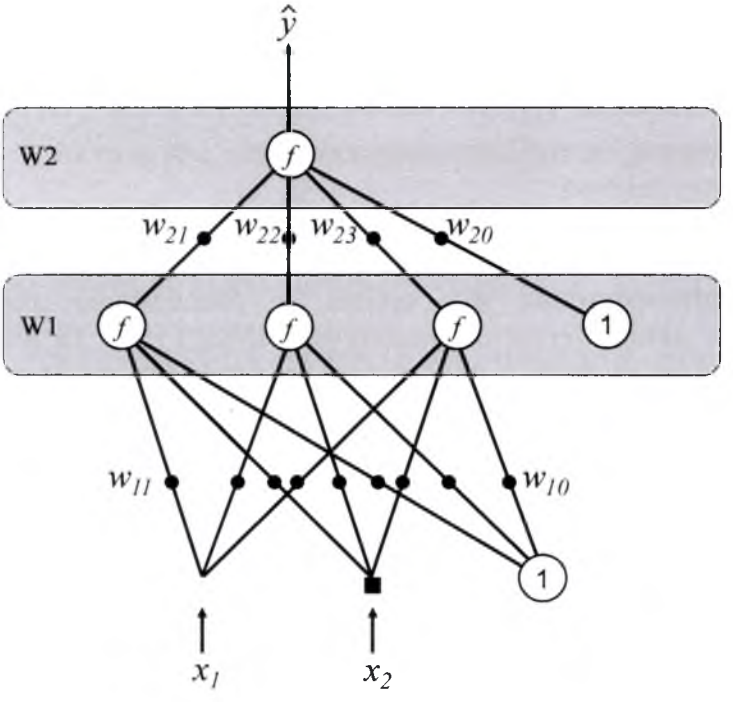
6.3 Struktura sieci

Jak wspomniano powyżej sieci można podzielić na jednowarstwowe i wielowarstwowe. Mianem warstwy określa się grupy neuronów, które nie są połączone między sobą, ale są połączone z neuronami innych warstw. Dodatkowo w neuronach tworzących warstwę musi zachodzić proces obliczeniowy [5]. Dlatego też węzły wejściowe formalnie nie tworzą warstwy neuronów. Każdy z typów sieci posiada wejście oraz warstwę wyjściową (Rys. 20 oraz Rys. 21).



Rys. 20 Przykładowy schemat jednowarstwowej jednokierunkowej sieci neuronowej skonstruowanej dla dwuwymiarowych danych wejściowych (x_1, x_2), sieć posiada jedno wyjście $\hat{y}$ w warstwie wyjściowej W1; z węzłem sieci (f) stowarzyszony jest wyraz wolny (1) oraz wagi w_i

Jeżeli sieć posiada więcej warstw to druga oraz każda kolejna warstwa nazywane są warstwami ukrytymi.



Rys. 21 Przykładowy schemat wielowarstwowej jednokierunkowej sieci neuronowej z jedną warstwą ukrytą (W1) skonstruowanej dla dwuwymiarowych danych wejściowych (x_1, x_2), sieć posiada jedno wyjście $\hat{y}$ w warstwie wyjściowej W2; z każdym węzłem sieci (f) stowarzyszony jest wyraz wolny (1) oraz wagi w_{ji}, gdzie j oznacza numer warstwy z której wychodzą wagi, a i to numer wagi w j-tej warstwie

Ilość warstw w sieci oraz węzłów przypadających na każdą z nich zależy od analizowanego problemu i wymaga optymalizacji. Sieci neuronowe z jedną warstwą ukrytą zawierającą kilka węzłów wyposażonych w nieliniowe funkcje aktywacji określane są mianem uniwersalnego aproksymatora (z ang. *universal approximator*). Oznacza to, iż sieć o takiej strukturze jest zdolna do aproksymacji każdej funkcji [54].

6.4 Uczenie sieci

Uczenie sieci jest procesem iteracyjnym. Polega ono na takim modyfikowaniu wag w_{ji}, aby zminimalizować błąd przewidywania zadanego wektora zmiennej zależnej. Istnieje wiele różnych definicji błędu. Do najbardziej rozpowszechnionych należą procent poprawnie sklasyfikowanych próbek w przypadku klasyfikacji i dyskryminacji oraz pierwiastek średniego błędu kwadratowego przewidywania dla próbek z niezależnego zbioru testowego dla kalibracji.

Procent poprawnie sklasyfikowanych próbek (z ang. *correct classification rate*, CCR) określa sumaryczną liczbę próbek przypisaną poprawnie do każdej z grup (Rów. 27) [55].

$$\mathrm{CCR}(f) = \frac{\sum_{i=1}^{m} \left| y_i + \hat{y}_i(f) \right|}{2m}; \qquad 27$$

gdzie: y to zmienna zależna, $\hat{y}_i(f)$ to wartość przewidziana dla y_i na podstawie modelu o określonej strukturze, m to liczba obiektów, wzór jest poprawny dla bipolarnego kodowania zmiennej zależnej.

Natomiast w kalibracji stosowany jest pierwiastek średniego błędu kwadratowego przewidywania dla próbek z niezależnego zbioru testowego (z ang. *root mean square terror of prediction*, RSMEP) [55]. Ta miara błędu mówi o mocy predykcyjnej modelu, a więc o jego zdolnościach przewidywania wartości zmiennej zależnej dla próbek, które nie brały udziału w konstrukcji modelu (Rów. 28).

$$\mathrm{RMSEP}(f) = \sqrt{\frac{\sum \left(y_i - \hat{y}_i(f) \right)^2}{m}}; \qquad 28$$

gdzie: y_i to i-ty element wektora zmiennej zależnej dla zbioru testowego, $\hat{y}_i(f)$ to wartość przewidziana dla y_i na podstawie modelu o dane strukturze, natomiast m to liczba elementów w niezależnym zbiorze testowym.

Obie miary błędu mogą być obliczane nie tylko dla niezależnego zbioru testowego, ale także dla zbioru modelowego i monitoringowego. Celem uniknięcia kolizji oznaczeń w niniejszej pracy przyjęto następujący zapis skrótów określający poszczególne błędy:

RMSE – pierwiastek średniego błędu kwadratowego przewidywania dla próbek ze zbioru modelowego charakteryzujący dopasowanie modelu do danych;
RMSEM – pierwiastek średniego błędu kwadratowego przewidywania dla próbek ze zbioru monitoringowego pozwalający określić optymalną kompleksowość lub architekturę modelu;
RMSEP – pierwiastek średniego błędu kwadratowego przewidywania dla próbek z niezależnego zbioru testowego charakteryzujący moc predykcyjną modelu;
CCR – procent poprawnie sklasyfikowanych próbek należących do zbioru modelowego charakteryzujący dopasowanie modelu do danych;
CCRM – procent poprawnie sklasyfikowanych próbek należących do zbioru monitoringowego pozwalający określić optymalną kompleksowość lub architekturę modelu;
CCRT – procent poprawnie sklasyfikowanych próbek należących do niezależnego zbioru testowego charakteryzujący moc predykcyjną modelu.

Pierwszym krokiem w uczeniu sieci neuronowej jest etap inicjalizacji wag. Istnieją różne podejścia do tego zagadnienia. Najbardziej powszechnym sposobem jest losowa inicjalizacja wag. W praktyce stosuje się wagi z przedziału <-1,1>, które zapewniają relatywnie dobrą generalizację i aproksymację. Inne metody inicjalizacji wag opisano w [5]:

Mając określoną architekturę sieci oraz zdefiniowane wstępne wartości wag, można przystąpić do procesu uczenia sieci. Polega on na takim modyfikowaniu połączeń pomiędzy neuronami, czyli zmianie wartości wag, aby jak najlepiej przewidzieć wartości zadanego wektora zmiennej zależnej. Poniżej przedstawiono najbardziej popularne podejście do tego zagadnienia: algorytm wstecznej propagacji błędu. Istnieją także jego modyfikacje jak np. podejście oparte na rekurencyjnej metodzie najmniejszych kwadratów [56].

6.4.1 Algorytm wstecznej propagacji błędu

Algorytm wstecznej propagacji błędu (z ang. *backpropagation*) jest podstawową metodą uczenia sieci neuronowych [57]. Metoda ta opracowana w roku 1974 przez P.J. Werbosa należy do grupy metod gradientowych. Nazwa tej metody nawiązuje do faktu, iż modyfikacja wag następuje w odwrotnej kolejności do kierunku przesyłania sygnału w sieci. Algorytm wstecznej propagacji błędu wykorzystuje pochodne funkcji aktywacji celem modyfikacji wag w warstwach ukrytych sieci.

Algorytm wstecznej propagacji błędu dla sieci o L warstwach, gdzie k=1:L, można zapisać następująco [6]:

$$y_i^{(k)}(n) = f(s_i^{(k)}(n)) ; \tag{29}$$

$$s_i^{(k)}(n) = \sum_{j=0}^{N_{k-1}} w_{ij}^{(k)}(n) x_j^{(k)}(n) ; \tag{30}$$

$$\varepsilon_i^{(k)}(n) = \begin{cases} d_i^{(L)}(n) - y_i^{(L)}(n), & k = L \\ \sum_{m=1}^{N_{k+1}} \delta_m^{(k+1)}(n) w_{mi}^{(k+1)}(n), & k = 1,\ldots,L-1 \end{cases}; \quad 31$$

$$\delta_i^{(k)}(n) = \varepsilon_i^{(k)}(n) f'(s_i^{(k)}(n)); \quad 32$$

$$w_{ij}^{(k)}(n+1) = w_{ij}^{(k)}(n) + 2\eta\delta_i^{(k)}(n) x_j^{(k)}(n); \quad 33$$

gdzie: n oznacza krok iteracyjny, $x_i^{(k)}$ oraz $y_i^{(k)}$ to odpowiednio i-te elementy sygnału wejściowego oraz wyjściowego neuronu; f to oznaczenie funkcji aktywacji neuronu; $d_i^{(k)}$ to i-ty element sygnału wzorcowego – zadana zmienna zależna; N_k to ilość węzłów w k-tej warstwie; $w^{(k)}$ to wagi w k-tej warstwie, $\varepsilon_i^{(k)}$ to błąd obliczony dla i-tego neuronu w k-tej warstwie; $s_i^{(k)}$ oznacza ważoną sumę wejść dla i-tego neuronu w k-tej warstwie; η to tzw. współczynnik uczenia określający stopień modyfikacji wag ($\eta > 0$).

Dzisiaj stosując algorytm wstecznej propagacji błędu używa się zmodyfikowanego równania 33 poprzez dodanie tzw. członu momentum (Rów. 34). Momentum wpływa na szybsze uzbieżnienie algorytmu.

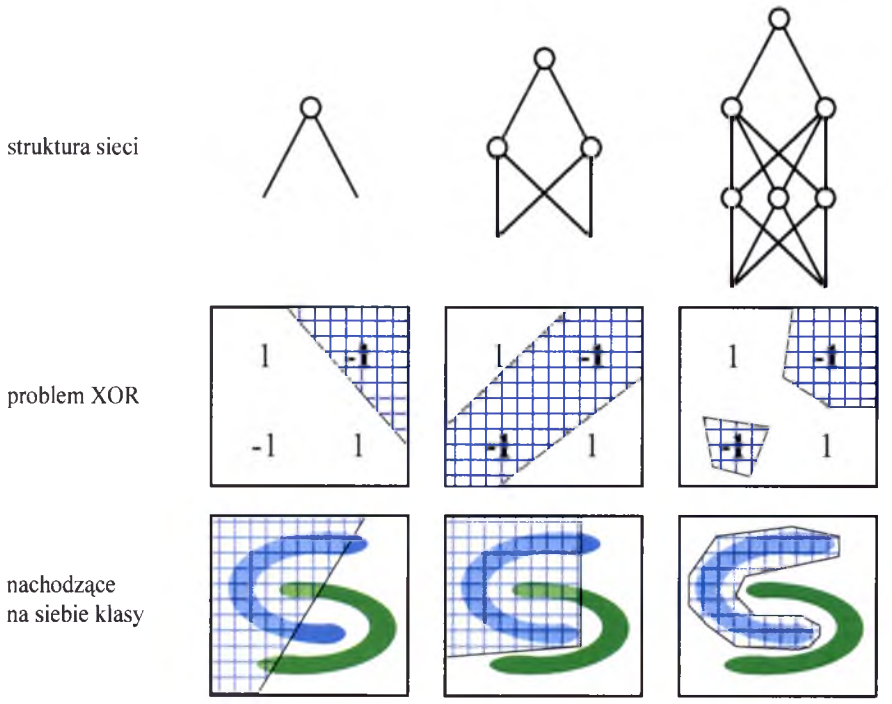
$$+\alpha\left[w_{ij}^{(k)}(n) - w_{ij}^{(k)}(n-1)\right]; \quad 34$$

gdzie: $\alpha \in \langle 0,1 \rangle$.

6.5 Optymalizacja architektury sieci

Jak już wspomniano wcześniej sieci neuronowe to systemy przetwarzające informacje, które posiadają zdolność aproksymacji praktycznie każdej funkcji matematycznej. Jedynym warunkiem koniecznym do pomyślnego wypełnienia tego zadania jest optymalna architektura sieci. Architekturę sieci neuronowej optymalizuje się dokonując niewielkich zmian w strukturze sieci. Podejście to, co prawda jest czasochłonne, ale w rękach doświadczonego analityka zapewnia bardzo dobre rezultaty.

Realizując zadanie kalibracji konstruuje się sieć wyposażoną w jedną warstwę ukrytą. Węzły warstwy ukrytej wyposażone są w nieliniowe funkcje aktywacji, natomiast w warstwie wyjściowej znajduje się liniowa funkcja aktywacji. Gdy sieć konstruowana jest celem klasyfikacji obiektów stosuje się nieliniowe funkcje aktywacji we wszystkich warstwach. Ilość warstw ukrytych zależy od rozkładu obiektów w przestrzeni modelowanych danych (Rys. 22).



Rys. 22 Zależność struktury sieci neuronowej od rozkładu obiektów w przestrzeni pomiarowej [48]

Optymalizacja architektury sieci neuronowej wymaga także odpowiedniego przygotowania danych [58]. Dostępne próbki dzieli się na trzy zbiory: zbiór modelowy (zwany także treningowym), zbiór monitoringowy oraz zbiór testowy. Do tego zadania najczęściej wykorzystuje się algorytm Kennarda i Stone'a [59, 60] lub algorytm Duplex [61] (szczegółowo opisane w rozdziale zatytułowanym *8 Modelowanie danych chemicznych*). Ilość obiektów w każdym z tych zbiorów jest różna i związana z jego funkcją. Zbiór modelowy używany do uczenia sieci tworzony jest tak, aby zawierał reprezentatywne próbki, a więc pochodzące z całej przestrzeni pomiarowej. Zbiór monitoringowy używany jest do określenia końca procesu uczenia sieci celem uniknięcia zjawiska przeuczenia sieci (z ang. *overfitting*). Zbiór testowy jest używany do walidacji skonstruowanego modelu. Zawiera on próbki, które nie brały udziału w konstrukcji modelu, dlatego jest nazywany niezależnym zbiorem testowym.

6.6 Zastosowania sieci neuronowych

Sieci neuronowe znalazły wiele zastosowań zarówno w nauce jak i przemyśle. Tego rodzaju systemy przetwarzania informacji umożliwiają wykonywanie zarówno działań klasyfikacyjnych jak i kalibracyjnych [58], sterowanie, filtrację, czy asocjację [62]. Do najpowszechniejszych zastosowań sieci neuronowych należą: diagnostyka systemów elektronicznych; optymalizacja różnych eksperymentów, sterowanie procesami i liniami produkcyjnymi w fabrykach; wszelkiego rodzaju prognozowanie, np. w ekonomi, medycynie, farmacji i meteorologii.

6.7 Wady i zalety sieci neuronowych

Zaletą sieci neuronowych jest ich zdolność do uczenia się w oparciu o przykłady. Zaletą jest także równoległy sposób przetwarzania informacji przez sieć, co znacznie przyspiesza obliczenia.

Interpretacja modelu sieci neuronowych jest utrudniona z uwagi na jego formę. Model sieci neuronowych jest bowiem rozproszony w zestawie wielu wag. Opis interpretacji modelu sieci neuronowych można znaleźć w [63]. Ta własność sieci neuronowych jest tym większym ograniczeniem ich stosowania, jeśli interpretacji konstruowanego modelu miałaby dokonać osoba nieposiadająca specjalistycznej wiedzy czy doświadczenia w danej dziedzinie.

7 Neuronowe systemy rozmyte

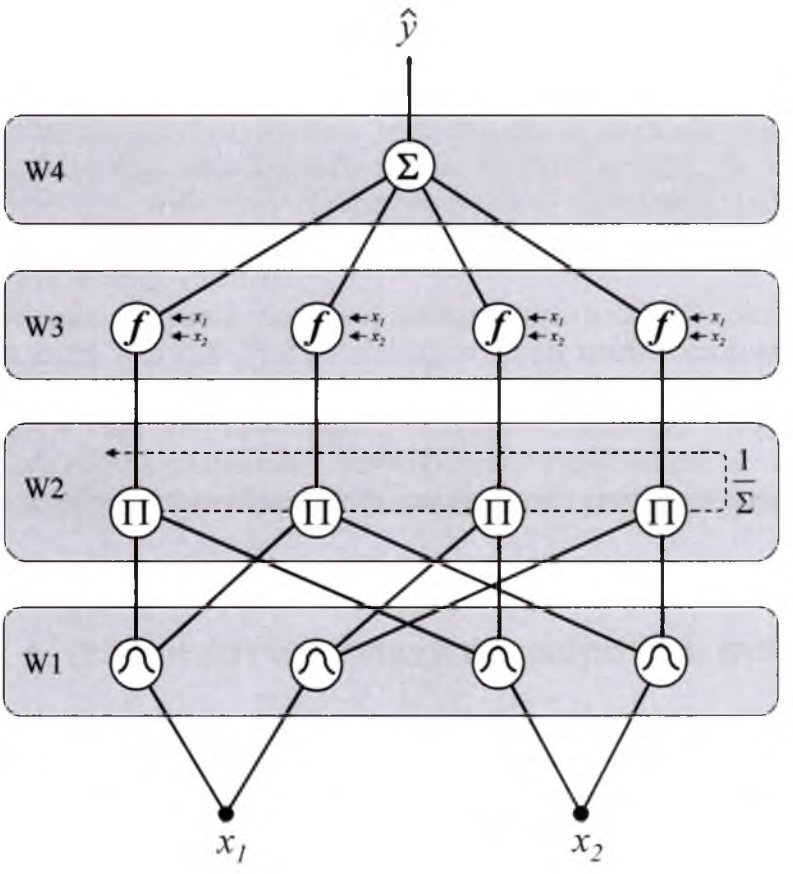
Połączenie systemów wnioskowania rozmytego i sieci neuronowych nosi nazwę neuronowych systemów rozmytych, w skrócie NFS [7, 8, 9]. W ramach modelu NFS, systemy wnioskowania rozmytego dostarczają schematu wnioskowania, a więc sposobu konstrukcji reguł logicznych, których uczenie odbywa się według algorytmu zaczerpniętego z teorii sieci neuronowych. Konstrukcja modelu NFS polega na automatycznym tworzeniu reguł logicznych na podstawie dostępnych danych.

7.1 Struktura neuronowych systemów rozmytych

NFS mają budowę warstwową, podobnie jak sieci neuronowe. Struktura modelu neuronowych systemów rozmytych nie jest sztywną konstrukcją, a zależy od analizowanego problemu. Podobnie jak w przypadku systemów wnioskowania rozmytego, istnieją różne typy neuronowych systemów rozmytych. W ramach neuronowych systemów rozmytych można, więc wnioskować w oparciu miedzy innymi o model Mamdani, lub Tsukamoto. Na rysunku 23 przedstawiono najbardziej popularny wariant neuronowych systemów rozmytych, mianowicie neuronowy system rozmyty typu Takagi, Sugeno i Kang (TSK-NFS) [7]. Neuronowy system rozmyty tego typu składa się z czterech warstw: warstwy wejściowej (W1), dwóch warstw ukrytych (W2 i W3) oraz warstwy wyjściowej (W4). Droga sygnału przez prezentowany na rysunku 23 schemat jest następująca:

W1 – Próbka w postaci wektora ($\mathbf{x} = [x_i, i = 1{:}n]$, gdzie n to liczba parametrów) jest podawana na pierwszą warstwę systemu, gdzie znajdują się węzły z nieliniowymi funkcjami przynależności. Następuje obliczenie wartości odpowiedzi funkcji przynależności. Ilość funkcji przynależności przypadających na każdy z parametrów (x_i) jest zależna od analizowanego problemu. Każda z takich funkcji przynależności jest opisana przez zestaw parametrów, np. parametry charakteryzujące funkcję Gaussa (Rys. 7a) to położenie maksimum piku, c_i, oraz jego szerokość, σ_i. Opisując elementy struktury neuronowego systemu rozmytego pojęcia funkcji przynależności oraz funkcji aktywacji można traktować zamiennie. Określenia parametrów funkcji przynależności dokonuje się przed przystąpieniem do konstrukcji modelu. W tym celu stosuje się różne techniki identyfikacji struktur (z ang. *structure identification*) oraz dzielenia przestrzeni (z ang. *space partitioning*) [8]. Wstępnie określone parametry funkcji przynależności są następnie poddawane modyfikacji w iteracyjnym procesie uczenia neuronowego systemu wnioskowania. Połączenia pomiędzy pierwszą oraz drugą warstwą są realizowane w taki sposób, aby zapewnić powstanie wszystkich możliwych

kombinacji funkcji przynależności pochodzących od każdego sygnału. Wyjściem z pierwszej warstwy są wartości przynależności próbek do podzbiorów, O1 (Rów. 6).



Rys. 23 Schemat neuronowego systemu rozmytego typu Takagi, Sugeno i Kang, gdzie: x_i to parametry wejściowe, $\hat{y}$ to przewidziana zmienna zależna, W1 to warstwa wejściowa, W2 oraz W3 to warstwy ukryte; w W2 wykonywane są kolejno operacje iloczynu (Π) i normalizacji $\left(\frac{1}{\Sigma}\right)$, a w W3 obliczana jest liniowa kombinacja oryginalnych zmiennych (f), W4 to warstwa wyjściowa gdzie następuje operacja sumowania wszystkich sygnałów (Σ); dla uproszczenia szaty graficznej nie zamieszczono węzłów reprezentujących wyrazy wolne

W2 – W drugiej warstwie następuje multiplikacja sygnałów z węzłów poprzedniej warstwy (Rów. 35).

$$O2 = w_i = \mu_i^{(A_i)}(x_1)\mu_i^{(B_i)}(x_2); \qquad 35$$

gdzie μ to funkcja przynależności próbki do zbioru A przypisana i-temu neuronowi.

Liczba węzłów tworzących drugą warstwę jest to liczba rozmytych reguł logicznych typu JEŻELI-TO. Jest ona równa liczbie funkcji przynależności przypadających na jedno wejście (x_i, i=1:n) podniesionej do potęgi równej liczbie wejść (n).

Przed przekazaniem sygnału do kolejnej warstwy każde wyjście z warstwy drugiej jest normalizowane (Rów. 36).

$$O2' = \overline{w_i} = \frac{w_i}{\sum_i w_i}; \qquad 36$$

W3 – Liczba węzłów tworzących trzecią warstwę jest równa liczbie węzłów w warstwie drugiej. W każdym węźle trzeciej warstwy oblicza się liniową kombinację oryginalnych zmiennych, $f_i(\mathbf{x})$ (Rów. 37) – neuronowy system rozmyty typu Takagi, Sugeno i Kang pierwszego rzędu.

$$O3 = f_i(\mathbf{x})\overline{w_i}; \qquad 37$$

Liniowa kombinacja oryginalnych zmiennych może być zastąpiona przez stałą wartość, a_i (Rów. 38). Mówi się wtedy o neuronowym systemie rozmytym Takagi, Sugeno i Kang zerowego rzędu.

$$O3 = a_i\overline{w_i}; \qquad 38$$

W4 - W tej warstwie znajduje się węzeł (lub węzły), w którym dokonywana jest operacja sumowania sygnałów z poprzedniej warstwy po przepuszczeniu ich przez funkcję przynależności znajdującą się w tym węźle, g_i (Rów. 39). Rodzaj użytej funkcji przynależności w warstwie czwartej zależy od zadania do jakiego dana sieć jest przeznaczona. Liniowa funkcja przynależności jest stosowana, gdy sieć realizuje zadanie kalibracji. W przypadku klasyfikacji w węźle znajduje się nieliniowa funkcja aktywacji (np. funkcja sigmoidalna, Rys. 6b).

$$O4 = \sum_i g_i O3; \qquad 39$$

W przypadku klasyfikacji liczba węzłów w warstwie wyjściowej zależy od liczby klas. Modelując problem dwuklasowy, wystarczy zastosować jeden węzeł. Jednakże dla problemów trzy- i więcej klasowego należy wyposażyć warstwę wyjściową sieci w tyle węzłów ile jest klas w modelowanych danych. Kiedy zadaniem sieci jest regresja, warstwę wyjściową będzie tworzył tylko jeden neuron. Na wyjściu z warstwy czwartej (O4) otrzymuje się wartość przewidzianą zmiennej zależnej dla modelowanej próbki. O4 jest końcowym wyjściem z neuronowego systemu rozmytego.

7.2 Uczenie neuronowego systemu rozmytego

Uczenie neuronowego systemu rozmytego może przebiegać według algorytmu wstecznej propagacji błędu [57, 64] i jest procesem iteracyjnym. Polega ono na dostosowywaniu parametrów funkcji przynależności znajdujących się w węzłach warstwy W1 i W4 (np. szerokość i położenie dla funkcji Gaussa), tak aby jak najlepiej przewidzieć wartość zmiennej zależnej dla badanej próbki. Mówi się wtedy

o dostrajaniu funkcji przynależności. Zmiany wag odbywają się w oparciu o tzw. uogólnioną regułę delta (z ang. *generalized delta rule*). Istnieją także inne algorytmy uczenia neuronowych systemów rozmytych, np. algorytm hybrydowy [7, 9].

Zgodnie z uogólnioną regułą delta obliczane są odpowiednie błędy dla warstwy wyjściowej ($B_{wyjścia}$, Rów. 40) oraz warstw ją poprzedzających ($B_{(i)}^{L}$, Rów. 41). Błędy warstw ukrytych $\left(B_{(i)}^{L}\right)$ są obliczane w oparciu o błędy neuronów warstwy wyższej $\left(B_{(i)}^{(L+1)}\right)$.

$$B_{wyjścia} = \sum_{t=1}^{k}\left(y_t - \hat{y}_t\right); \qquad 40$$

$$B_{(i)}^{L} = f'^{L}_{(i)} \sum B_{(k)}^{(L+1)} w_{(ki)}^{L+1}; \qquad 41$$

gdzie: L oznacza warstwę, i, k to indeksy węzłów odpowiednio w L-tej oraz (L+1)-tej warstwie, w_{ik} oznacza wagę pomiędzy i-tym węzłem w warstwie L oraz k-tym węzłem w warstwie L+1, $\hat{y}_t$ to y przewidziany.

Dla warstwy, w której przeprowadzana jest operacja iloczynu, błąd oblicza się według równania 42, a wagi modyfikuje zgodnie z równaniem 44.

$$B_{(i)}^{L} = f'^{L}_{(i)} \sum_{k}\sum B_{(k)}^{(L+1)}\left(\prod_{k} w_{(km)}^{(L+1)} \prod_{m=1} o_{(m)}^{L}\right); \qquad 42$$

Obliczone błędy służą do modyfikacji wag. Zmiany wag oblicza się według równania 43. Dla wag łączących warstwy W2 i W3 stosuje się równanie 44.

$$w_{(ik)}^{L}(t+1) = w_{(ik)}^{L}(t) \times \left(1 + B_{(i)}^{L}\right) w_{(ik)}^{L}(t) o_{(k)}^{(L-1)}; \qquad 43$$

$$w_{(ik)}^{L}(t+1) = w_{(ik)}^{L}(t) + \eta B_{(i)}^{L} o_{(k)}^{(L-1)}; \qquad 44$$

gdzie: t oznacza numer iteracji, η to współczynnik uczenia. Więcej szczegółów o współczynniku uczenia można znaleźć w [65].

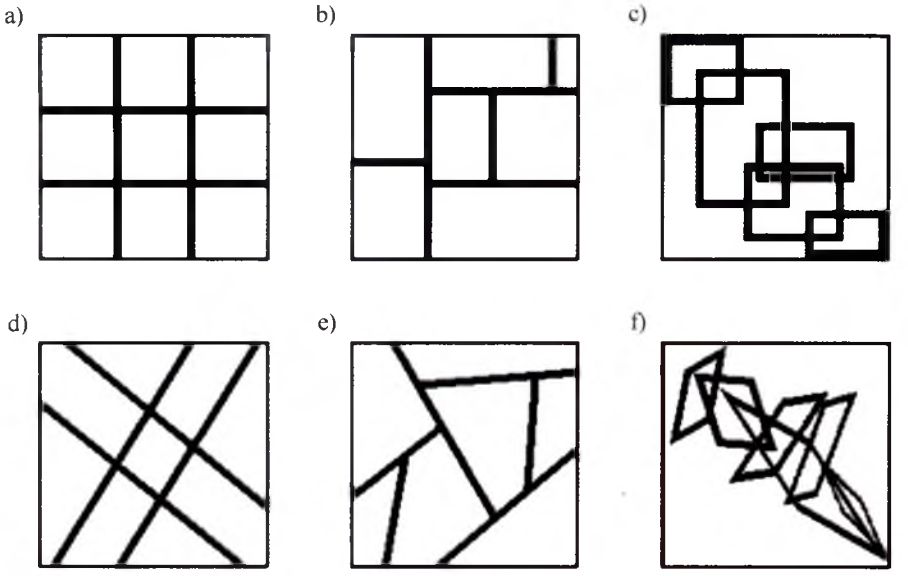
Modelując dane przy użyciu neuronowych systemów rozmytych można wybrać typ systemu: Mamdani; Takagi, Sugeno i Kang czy Tsukoamoto. Konstruując model należy zadecydować, ile funkcji przynależności zostanie użytych. Ponadto określenia wymaga zastosowana strategia uczenia modelu (np. metoda hybrydowa czy metoda wstecznej propagacji błędu). Wszystkie te czynniki sprawiają, iż konstruuje się tak naprawdę nie jeden model, a co najmniej kilkanaście modeli, spośród których należy wybrać ten najlepszy, tzn. ten o najlepszej mocy predykcyjnej i minimalnej kompleksowości. Kompleksowa strategia modelowania danych chemicznych omówiona zostanie w rozdziale zatytułowanym *8 Modelowanie danych chemicznych*.

7.3 Identyfikacja struktury danych oraz dzielenie przestrzeni

Etap rozpoznawania struktury danych jest ważną częścią modelowania przy użyciu neuronowych układów rozmytych [7]. Jest to pierwszy i wspólny krok dla wszystkich typów neuronowych systemów rozmytych. Etap ten polega na podzieleniu przestrzeni pomiarowej danych na podprzestrzenie. Podziału przestrzeni pomiarowej dokonuje się przez jej opis z wykorzystaniem funkcji przynależności. Określa się liczbę funkcji przynależności przypadającą na jedną zmienną (jedno wejście). Wyznacza się także położenie funkcji w przestrzeni pomiarowej, np. położenie maksimum piku dla funkcji Gaussa czy położenie punktu przegięcia dla funkcji sigmoidalnej. Najbardziej popularne typy podziału przestrzeni pomiarowej to podział kratkowy (z ang. *grid partition*), drzewkowy (z ang. *tree partition*) oraz rozproszony (z ang. *scatter partition*). Do stosowanych w tym celu metod należą metody grupowania np. drzewa klasyfikacji i regresji CART, algorytm C-środków (z ang. *K-means* [66]) oraz jego rozmyta modyfikacja [67].

7.3.1 Typy podziału przestrzeni pomiarowej

Kratkowy podział przestrzeni pomiarowej to najprostszy typ podziału przestrzeni, który używany jest wtedy, gdy przestrzeń pomiarowa jest definiowana tylko przez kilka parametrów (Rys. 24a, d) [7]. W miarę wzrostu liczby parametrów gwałtownie rośnie ilość reguł logicznych. Mówi się wtedy o tak zwanym przekleństwie wymiarowości (z ang. *curse of dimensionality*). Drzewkowy podział przestrzeni danych tworzy w tejże przestrzeni regiony, które nie są sobie równe (Rys. 24b, e).



Rys. 24 Podział przestrzeni według różnych typów: a, d) kartkowy; b, e) drzewkowy; c, f) rozproszony; na oryginalnych zmiennych (a-c) oraz na transformacjach zmiennych (d-f)

Drzewkowy podział przestrzeni pomiarowej ma na celu redukcję liczby funkcji przynależności, a więc i liczbę reguł logicznych. Przykładem metody dzielącej przestrzeń pomiarową w sposób drzewkowy są drzewa klasyfikacji i regresji, CART.

Kolejnym typem podziału przestrzeni danych jest podział rozproszony. W tym wypadku także są tworzone nierówne regiony w przestrzeni pomiarowej. Rozproszony podział przestrzeni pomiarowej pozwala na zmniejszenie liczby reguł logicznych do rozsądnej ilości eliminującej przekleństwo wymiarowości. Funkcje przynależności mogą być zdefiniowane bezpośrednio na zmiennych (Rys. 24c) lub na ich transformacji (Rys. 24f). Przykładem metody pozwalającej na taki podział przestrzeni danych jest metoda grupowania różnicowego [7, 67].

7.3.2 Fuzzy C-means

Z uwagi na temat pracy w tym podrozdziale opisana zostanie rozmyta modyfikacja algorytmu C-środków o nazwie *Fuzzy C-means* (FCM) [68] oraz wykorzystywana metoda grupowania różnicowego [69].

FCM to technika grupowania próbek, której algorytm może być zapisany w czterech krokach:

1. Każdą próbkę $\mathbf{x}_j$ przypisuje się do każdego klasteru za pomocą losowej wartości przynależności μ_{ij}. Robi się to w taki sposób, aby zsumowane wartości przynależności danej próbki do jednego klasteru były równe jedności.
2. Następnie obliczany jest centroid każdego klasteru $\mathbf{c}_i$ dla n zmiennych (Rów. 45), gdzie m to tzw. parametr rozmycia. Obliczane są także odległości euklidesowe d_{ij} wszystkich próbek od każdego centroidu (Rów. 46).

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m \mathbf{x}_j}{\sum_{j=1}^{n} \mu_{ij}^m}; \tag{45}$$

$$d_{ij} = \|\mathbf{c}_i - \mathbf{x}_j\|; \tag{46}$$

gdzie j to ilość próbek, i oznacza centroid klasteru.

3. Obliczana jest funkcja kosztów (Rów. 47). Jeżeli jej wartość jest poniżej zadanej wartości progowej algorytm zostaje przerwany, jeśli nie przechodzi się do kroku 4. Celem jest minimalizacja następującego kryterium:

$$J = \sum_{i=1}^{C} \sum_{j=1}^{N} \mu_{ij}^m d_{ij}^2; \tag{47}$$

gdzie: d_{ij} C to liczba próbek, N to liczba klasterów definiowana *a priori* przez konstruktora; m to parametr rozmycia. Procedura grupowania obiektów

jest tym bardziej rozmyta im parametr rozmycia jest większy. Szczegóły na temat algorytmów FCM można znaleźć w [70].

4. Obliczane są nowe wartości przynależności obiektów do klasterów zdefiniowanych przez centroidy $\boldsymbol{c}_i$ według poniższego równania.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}}; \qquad 48$$

Wszystkie kroki są powtarzane do uzyskania zbieżności.

7.3.3 Grupowanie różnicowe

Metoda grupowania różnicowego (z ang. *subtractive clustering*) [67] to metoda grupowania próbek bazująca na miarze gęstości otoczenia próbek. Na początku każda próbka jest potencjalnym centrum klasteru. Dla każdego obiektu $\mathbf{x}_i$, obliczana jest miara gęstości D_i według poniższego równania:

$$D_i = \sum_{j=1}^{n} exp\left(- \frac{\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2}{(r_a / 2)^2} \right); \qquad 49$$

gdzie; $\mathbf{x}_i$ to obiekt rozpatrywany jako centrum klasteru, $\mathbf{x}_j$ to obiekt w promieniu r_a od $\mathbf{x}_i$. Parametr r_a to dodatnia stała odzwierciedlająca promień sąsiedztwa, poza którym obiekty nie mają wpływu na wartość gęstości D_i.

Obiekt wybrany na centrum klasteru jako pierwszy ($\mathbf{x}_{C1}$) to ten, który cechuje się największą wartością miary gęstości. Następnie wartości gęstości dla pozostałych obiektów są przeliczane według równania 2.

$$D_i^{nowa} = D_i - D_{C_k} exp\left(- \frac{\left\| \mathbf{x}_i - \mathbf{x}_{C_k} \right\|^2}{(r_b / 2)^2} \right); \qquad 50$$

gdzie: D_{C_k} to wartość miary gęstości dla poprzednio wybranego centroidu, r_b to podobnie jak w r_a promień sąsiedztwa jakim obiekty powinny mieć zmniejszoną gęstość, zwykle $r_b > r_a$, przyjęło się stosować $r_b = 1{,}5 r_a$.
Jako centrum drugiego klasteru wybrany jest ponownie obiekt o największej wartości gęstości D_i^{nowa}. Algorytm kończy się w momencie osiągnięcia założonej liczby klasterów. Innym sposobem zakończenia procedury jest założenie:

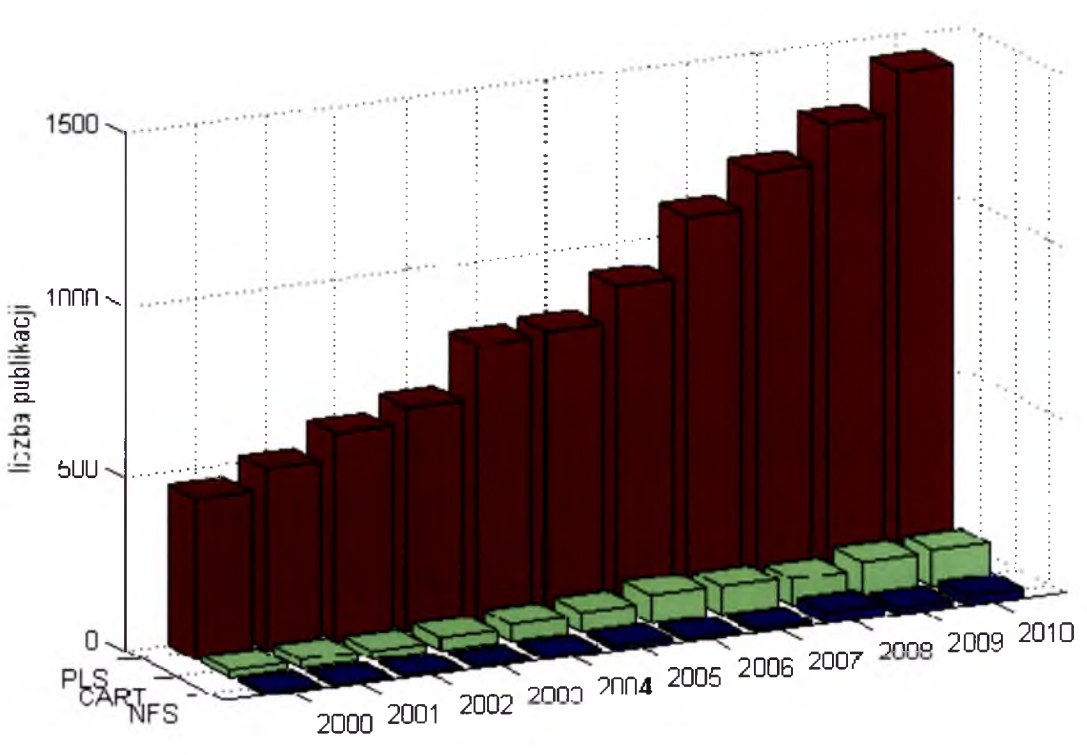
$D_k < \varepsilon D_{C_1}$; 51

gdzie: D_k to wartość gęstości dla k-tego klasteru; ε to parametr określający stosunek liczby klasterów do liczebności danych. Szczegóły dotyczące wyznaczania ε zawiera [71].

7.4 Zastosowania neuronowych systemów rozmytych - przegląd literaturowy

Neuronowe układy rozmyte mają wiele zastosowań w różnych dziedzinach nauki, takich jak np. medycyna [72,73], genetyka [74] i geologia [75]. Zastosowania NFS znajduje się miedzy innymi w ekonomi oraz planowaniu czy przewidywaniu w biznesie [76]. Także różne gałęzie przemysłu doceniły możliwości, jakie daje modelowanie przy użyciu NFS, należą do nich np. przemysł medyczny, farmaceutyczny oraz elektroniczny (RTv i AGD; [77]).

Dokonany przegląd literaturowy wykazał, iż w chemii zastosowania neuronowych układów rozmytych są niezwykle rzadkie [78].



Rys. 25 Zestawienie ilości publikacji, w których stosowano metody: PLS, CART i NFS na przestrzeni ostatniej dekady do modelowania danych chemicznych

Na powyższym wykresie (Rys. 25) przedstawiono liczbę publikacji na temat zastosowań NFS, drzew klasyfikacji i regresji oraz metody częściowych najmniejszych kwadratów w chemii na przestrzeni ostatnich 11 lat. Metoda częściowych najmniejszych kwadratów (PLS) [79, 80] jest najpowszechniej stosowaną liniową metodą modelowania i została pokazana na Rys. 25 w celach porównawczych.

Przeprowadzony przegląd literaturowy wykazał, iż Dotychczasowe nieliczne dzastosowania NFS w chemii [81,82,83,84] to:

- modelowanie QSAR [85,86,87,88];
- modelowanie określonych własności w oparciu o widma w NIR, czy NMR [89,90];
- w przemyśle petrochemicznym do analizy różnego rodzaju paliw w oparciu o dane uzyskane techniką chromatografii gazowej (GC-DMS) [91,92];
- oznaczanie składu pierwiastkowego próbek w oparciu o dane spektroskopowe fluorescencji promieni rentgenowskich [93];
- dyskryminacja pacjentów chorych na nowotwór w oparciu o dane proteomiczne oraz genomiczne [94,95];
- przewidywanie szeregów czasowych [76];
- analiza danych z brakującymi elementami [96];
- wybór zmiennych istotnych [97];
- kontrola procesów chemicznych [98].

7.5 Wady i zalety neuronowych systemów rozmytych

Zalety neuronowych układów rozmytych:

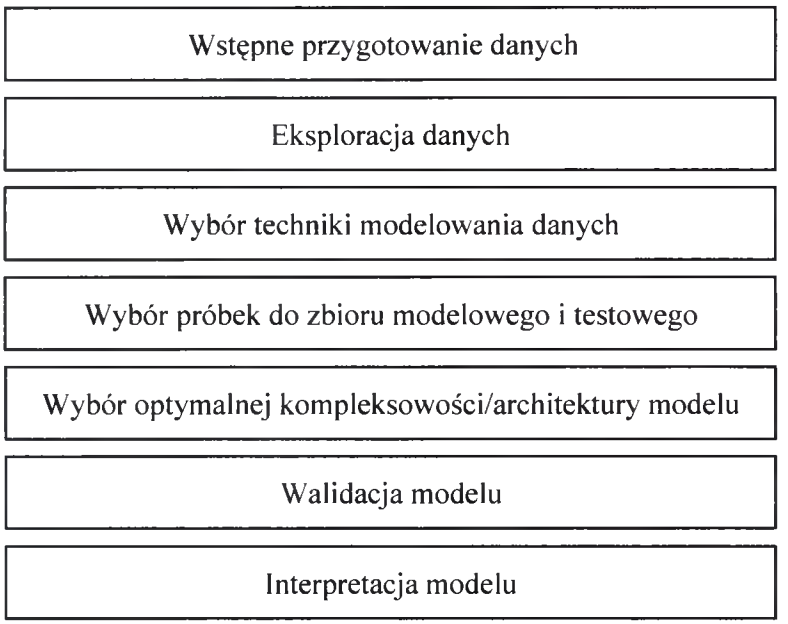
- automatyczna konstrukcja reguł logicznych;
- odporność na błędy, braki precyzji i jednoznaczności w danych;
- możliwość łatwej interpretacji modelu dzięki zmiennym lingwistycznym;
- eksploracja i uczenie się przestrzeni pomiarowej.

Wady neuronowych układów rozmytych:

- zależność od inicjalizacji parametrów funkcji przynależności;
- przekleństwo wymiarowości.

8 Modelowanie danych chemicznych

Wybór odpowiedniej techniki modelowania danych jest tylko jednym z kilku problemów, przed jakim staje analityk chcąc modelować dane. Proces modelowania danych wymaga odpowiedzi na szereg ważnych i niejednokrotnie trudnych pytań, od których zależeć będzie jakość skonstruowanych modeli. Ogólna strategia konstruowania modelu dla danych eksperymentalnych zakłada siedem podstawowych etapów (Rys. 26).



Rys. 26 Podstawowe etapy modelowania danych eksperymentalnych

8.1 Metody wstępnego przygotowania danych do analizy

Wstępne przygotowanie danych jest najważniejszym etapem modelowania determinującym jakość otrzymanych modeli. Zazwyczaj ma ono na celu eliminację i/lub korekcję niepożądanych efektów fizycznych w danych, co wpływa na poprawę interpretacji modelu [3, 4, 99]. Wśród istniejących metod wstępnego przygotowania danych do analizy można wyróżnić trzy grupy:

– Pierwsza grupa obejmuje metody mające na celu modyfikację indywidualnych zmiennych, a w skrajnych przypadkach ich eliminację. Do tej grupy metod należą między innymi centrowanie, autoskalowanie (zwane także standaryzacją) czy transformacja logarytmiczna.

– Drugą grupę stanowią metody modyfikujące całe próbki, czyli obiekty w danych, np. profile spektralne czy stężeniowe. Do metod tych zaliczyć można techniki normalizacyjne, pochodne, algorytmy eliminacji szumu czy linii bazowej.

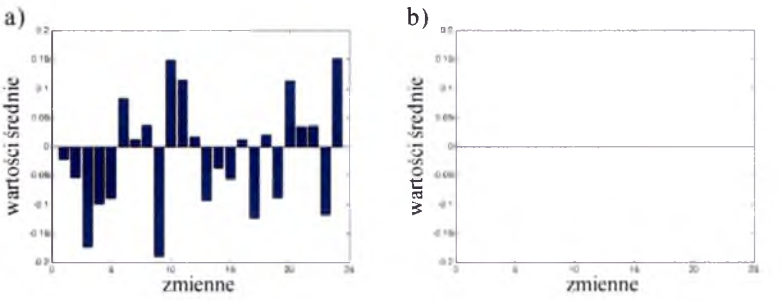
– Trzecią grupę metod (niewykorzystywanych w niniejszej pracy) stanowią metody nakładania sygnałów instrumentalnych.

8.1.1 Centrowanie

Jedną z najczęściej stosowanych metod wstępnego przygotowania danych do analizy jest centrowanie [3]. Ma ono na celu usunięcie z danych tej części informacji, która nie wpływa na zróżnicowanie próbek. Cel zostaje osiągnięty przez odjęcie od każdego elementu parametru średniej tego parametru (Rów. 52). Efektem tej transformacji jest przesunięcie środka danych do początku kartezjańskiego układu współrzędnych. Średnia każdego parametru po centrowaniu wynosi zero (Rys. 27).

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}^T \bar{\mathbf{x}}; \qquad 52$$

gdzie: $\mathbf{X}_c$ to wycentrowana macierz danych $\mathbf{X}$, $\mathbf{1}^T$ to kolumnowy wektor jednostkowy, $\bar{\mathbf{x}}$ to wektor zawierający średnie z macierzy $\mathbf{X}$ obliczane po kolumnach.



Rys. 27 Średnie dla 25 parametrów symulowanej macierzy danych a) przed oraz b) po centrowaniu

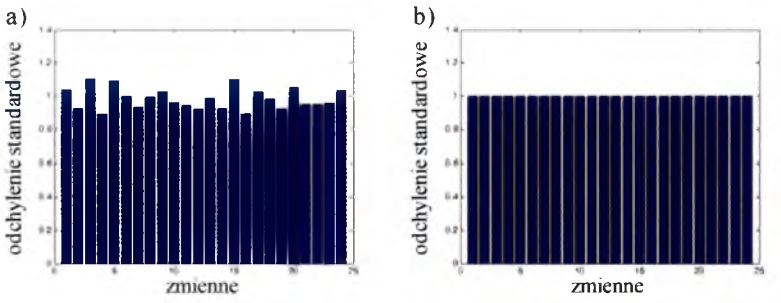
8.1.2 Standaryzacja

W chemii często pracuje się z danymi zawierającymi parametry wyrażone w różnych jednostkach czy zakresach [3]. Przykładem takich danych mogą być dane dotyczące oznaczania własności próbek wody: np. $[F^-]$ w μg/L, $[Ca^{2+}]$ w mg/L oraz pH. Zastosowanie procedury standaryzacji danych pozwala na zniwelowanie ewentualnego wpływu spowodowanego różnicami zakresów przez nadanie wszystkim mierzonym

parametrom jednakowej istotności. Matematycznie odbywa się to według równania 53 i oznacza nadanie wszystkim parametrom jednostkowego odchylenia standardowego (Rys. 28).

$$\mathbf{X}_s = \mathbf{X}./\left(\mathbf{1}^T \boldsymbol{\sigma}\right); \qquad 53$$

gdzie: $\mathbf{X}_s$ to wystandaryzowana macierz danych $\mathbf{X}$, $\mathbf{1}^T$ to kolumnowy wektor jednostkowy, σ to wektor zawierający odchylenia standardowe z macierzy $\mathbf{X}$ obliczane po kolumnach, natomiast ./ oznacza operacje ilorazu odpowiadających sobie elementów macierzy



Rys. 28 Odchylenia standardowe dla 25 parametrów symulowanej macierzy danych a) przed oraz b) po standaryzacji

8.1.3 Autoskalowanie

Połączenie centrowania i standaryzacji nosi nazwę autoskalowania [3]. W skutek zastosowania tejże procedury, uzyskujemy dane o rozkładzie zbliżonym do normalnego.

$$\mathbf{X}_a = \left(\mathbf{X} - \mathbf{1}^T \bar{\mathbf{x}}\right)./\left(\mathbf{1}^T \boldsymbol{\sigma}\right); \qquad 54$$

gdzie znaczenie symboli jest analogiczne do poprzednich równań.

8.1.4 Transformacja SNV

Transformacja SNV (z ang. *standard normal variate*) jest często stosowana do wstępnego przygotowywania sygnałów instrumentalnych takich jak widma w bliskiej podczerwieni czy chromatogramy. Dokonując transformacji SNV centruje się wiersze macierzy danych ich wartościami średnimi, a następnie standaryzuje się odpowiednimi odchyleniami standardowymi. Transformacja SNV wykonywana jest według równania 54 na transponowanej macierzy $\mathbf{X}$. W wyniku tej operacji każdy z sygnałów (wierszy) ma jednostkową wariancję.

8.2 Eksploracja danych oraz wybór techniki modelowania

Drugim etapem modelowania jest eksploracja danych pozwalająca na badanie rozkładu obiektów w przestrzeni pomiarowej. Powszechnie stosowanymi metodami eksploracji i wizualizacji danych są analiza czynników głównych, PCA [100] i hierarchiczne metody grupowania danych (CA) [67]. Elastyczny odpowiednik analizy czynników głównych (rPCA) pozwala na lepszą identyfikację obiektów odległych w przestrzeni danych [101, 102].

Kolejnym, trzecim krokiem jest wybór techniki modelowania danych. W zależności od charakteru modelowanych danych można użyć liniową (np. MLR [3], PCR [103], PLS) lub nieliniową metodę (np. CART, ANN, RBFN [7]).

8.3 Podział obiektów na zbiory

Czwartym etapem konstrukcji modelu jest podział dostępnych próbek (obiektów) na zbiory, z których jeden służy do konstrukcji modelu, drugi do określenia jego optymalnej architektury czy kompleksowości modelu, a trzeci do sprawdzenia jego mocy przewidywania, czyli walidacji. Zbiory te nazywane są odpowiednio zbiorem modelowym (lub treningowym), monitoringowym oraz testowym. W przypadku sieci neuronowych czy neuronowych układów rozmytych zbiór monitoringowy służy do określenia lokalnego minimum, a więc momentu zaprzestania uczenia sieci. (Alternatywą dla konstrukcji zbioru monitoringowego może być procedura walidacji krzyżowej różnego typu.) Ostatni konstruowany zbiór nie bierze udziału w konstrukcji modelu, dlatego nazywany jest niezależnym zbiorem testowym. Zbiór testowy wykorzystywany jest jedynie do oceny mocy predykcyjnej skonstruowanego modelu (do walidacji). Obiekty zwykle dzieli się na zbiór modelowy, monitoringowy i testowy w stosunku 60% : 20% : 20%, jednak proporcje te nie są arbitralne i zależą od liczebności modelowanych danych.

Powszechnie stosuje się dwie metody podziału próbek na zbiory: algorytm Kennarda & Stone'a [59, 60] oraz algorytm Duplex [61]. Metody te różnią się sposobem konstrukcji zbiorów, a więc nawet, jeśli zbiory będą tej samej liczebności mogą one, w mniejszej lub większej części, zawierać różne obiekty. Obie metody dzielą obiekty na dwa zbiory: zbiór modelowy oraz testowy. Aby podzielić dane na trzy zbiory należy zbiór testowy ponownie podzielić na dwa zbiory: zbiór monitoringowy oraz właściwy zbiór testowy. Takie podejście zostało wykorzystane w niniejszej pracy do obróbki danych (patrz rozdział *9 Analizowane dane i wyniki*).

8.3.1 Algorytm Kennarda i Stone'a

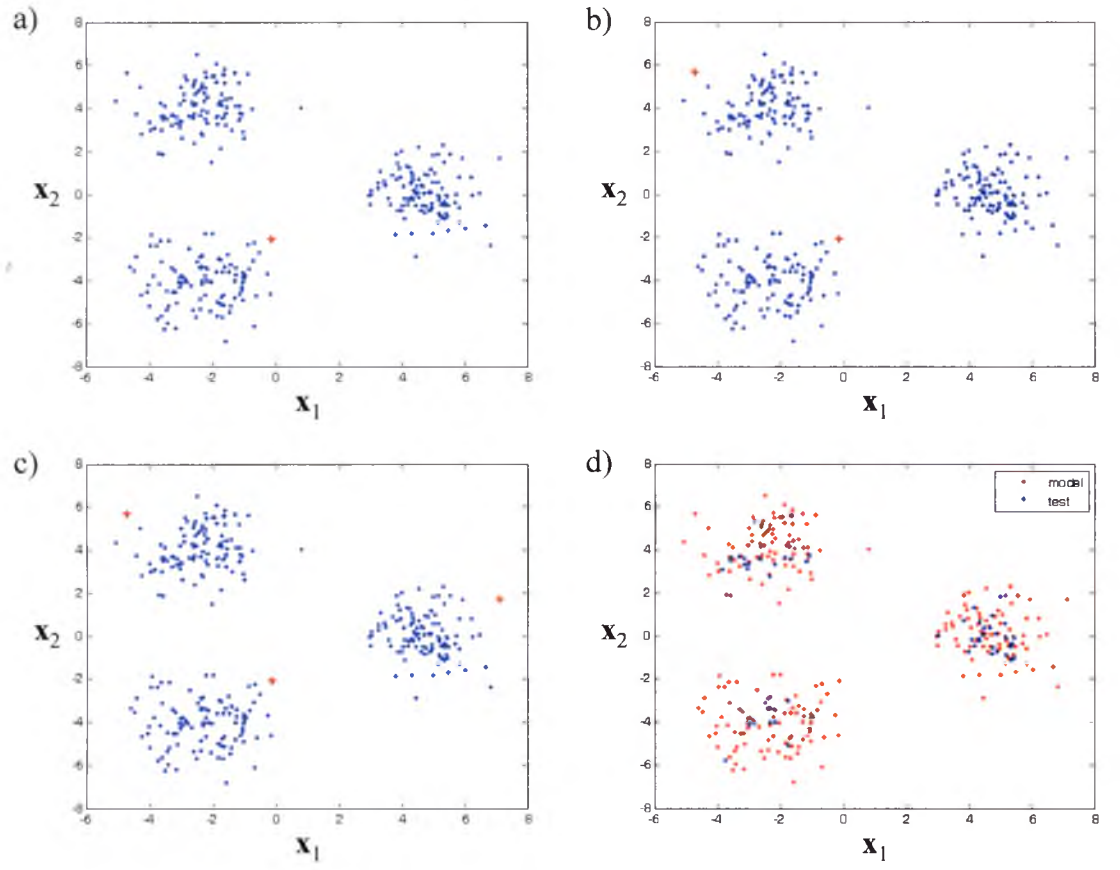
Algorytm Kennarda i Stone'a umożliwia wybór reprezentatywnego zbioru modelowego. W tym celu, z przestrzeni danych wybierane są obiekty jak najbardziej różne od siebie. Uzyskuje się w ten sposób podzbiór relatywnie równomiernie pokrywający oryginalny zbiór danych. O różnicy obiektów wnioskuje się wykorzystując kryterium ich podobieństwa, jakim jest odległość euklidesowa:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}; \qquad 55$$

gdzie: d_{ij} to odległość euklidesowa pomiędzy dwoma wektorami ($\mathbf{x}_i$ oraz $\mathbf{x}_k$) opisującymi i-tą oraz j-tą próbkę w n-wymiarowej przestrzeni.

Celem zapewnienia reprezentatywności zbioru modelowego w pierwszej kolejności wybierane są obiekty właśnie do tego zbioru. Wybór próbek odbywa się według następujących kroków:

- pierwszy wybrany obiekt to ten najbliżej średniej (Rys. 29a);
- drugi wybrany obiekt znajduje się najdalej od pierwszego (Rys. 29b);
- każdy kolejny obiekt wybierany jest w taki sposób, aby był położony jak najdalej od już wybranych próbek (Rys. 29c, d).



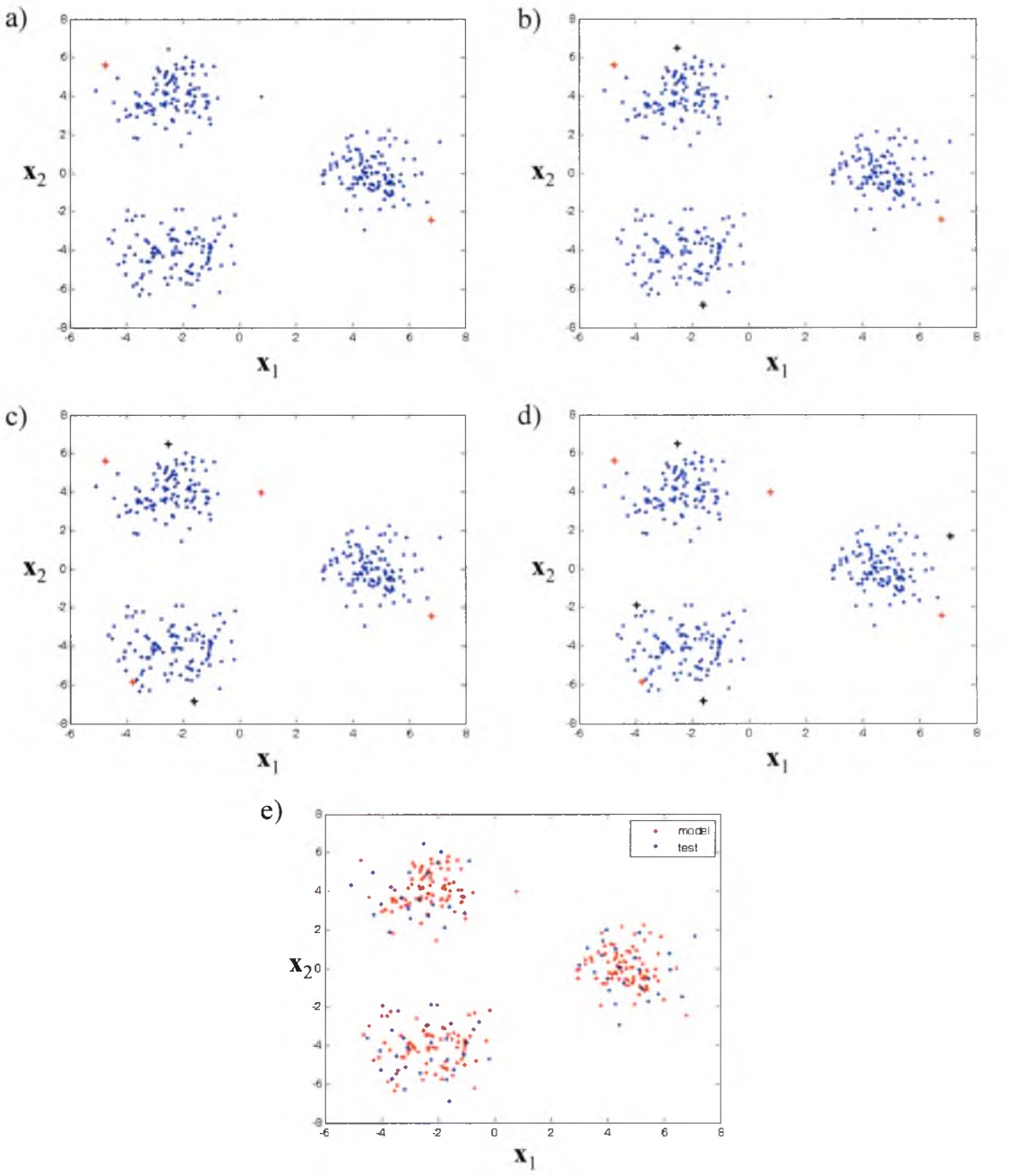
Rys. 29 a-c) Kolejność wyboru obiektów do zbioru modelowego (+) za pomocą algorytmu Kennarda i Stone'a dla dwuwymiarowych symulowanych danych ($\mathbf{x}_1$, $\mathbf{x}_2$) zawierających 300 obiektów tworzących trzy równoliczne grupy oraz d) rezultat zastosowanej procedury gdzie podzielono obiekty w stosunku 75% do zbioru modelowego i 25% do zbioru testowego

Koniec algorytmu następuje wtedy, gdy w zbiorze modelowym znajdzie się wcześniej założona liczba obiektów. Matematyczne kryterium wyboru obiektów przedstawia poniższe równanie:

$$\max_{i_0}\left\{\min_{i}\left(d_{1i}, d_{2i}, \ldots, d_{i_0 i}\right)\right\}; \qquad 56$$

gdzie: i_0 oraz i oznacza obiekty będące odpowiednio kandydatami do zbioru modelowego oraz będące w zbiorze modelowym.

8.3.2 Algorytm Duplex



Rys. 30 a-d) Kolejność wyboru obiektów do zbioru modelowego (+) oraz testowego (+) za pomocą algorytmu Duplex dla dwuwymiarowych symulowanych danych ($\mathbf{x}_1$, $\mathbf{x}_2$) zawierających 300 obiektów tworzących trzy równoliczne grupy oraz e) rezultat zastosowanej procedury gdzie podzielono obiekty w stosunku 75% do zbioru modelowego i 25% do zbioru testowego

Algorytm Duplex jest metodą wybory próbek zapewniającą reprezentatywność zarówno zbioru modelowego jak i testowego. Podobnie jak w przypadku algorytmu Kennarda i Stone'a wykorzystuje się tutaj odległość euklidesową jako miarę podobieństwa pomiędzy obiektami. Jednakże obie metody różnią się sposobem wyboru obiektów do tworzonych podzbiorów. Jako pierwsze do zbioru modelowego trafiają dwa najbardziej oddalone od siebie obiekty (Rys. 30a). Następne dwa najbardziej oddalone od siebie obiekty trafiają do zbioru testowego (Rys. 30b). Kolejne obiekty trafiają naprzemiennie do obu zbiorów (Rys. 30c-e). Procedura zostaje zakończona, gdy zbiór modelowy osiągnie założoną wcześniej liczebność.

8.4 Kompleksowość, walidacja oraz interpretacja modelu

Kolejnym etapem konstrukcji modelu jest określenie jego kompleksowości lub struktury. O kompleksowości modelu mówi się w przypadku takich metod jak PLS czy PCR, gdzie określenia wymaga optymalna liczba zmiennych ukrytych użytych do konstrukcji modelu. Z kolei o optymalnej architekturze mówi się w przypadku takich metod jak CART czy ANN. Optymalizacja architektury modelu to taki dobór jego parametrów (liczba węzłów czy warstw), który zapewnia najlepsze wyniki. Najpowszechniej stosowaną metodą określania kompleksowości modelu czy jego struktury jest walidacja krzyżowa różnego typu, np. wyrzuć-*n*-obiektów albo typu Monte-Carlo [55]. Innym sposobem jest użycie zbioru monitoringowego. Jednak wybór optymalnej kompleksowości modelu czy jego architektury nie jest prosty i niejednokrotnie wymaga skonstruowania wielu modeli.

Walidację skonstruowanego modelu przeprowadza się z wykorzystaniem niezależnego zbioru testowego. To właśnie ten zbiór pozwala na oszacowanie mocy predykcyjnej modelu. Dla modeli kalibracyjnych obliczany jest pierwiastek średniego błędu kwadratowego dla próbek z niezależnego zbioru testowego (RMSEP, Rów. 28). Natomiast dla modeli dyskryminacyjnych stosuje się miarę błędu będącą procentem poprawnie sklasyfikowanych próbek z niezależnego zbioru testowego (CCRt, Rów. 27).

Ostatnim krokiem jest interpretacja modelu. Krok ten nie jest obligatoryjny i nie został omówiony w niniejszej pracy. Istnieją takie metody jak CART czy NFS, które stosuje się właśnie z myślą o interpretacji modelu. Z kolei np. w przypadku sieci neuronowych interpretacja modelu zwykle nie jest stosowana.

9 Analizowane dane i wyniki

Przedstawione w poprzednich rozdziałach założenia teoretyczne dotyczące procesu modelowania danych chemicznych zostały wykorzystane w praktyce. Poniżej zamieszczono opis modelowanych danych chemicznych oraz wyniki przeprowadzonych analiz. Wykorzystane dane mierzono zarówno za pomocą technik instrumentalnych (np. dane spektroskopowe) jak i tradycyjnych metod analitycznych. Wyboru przedstawionych danych dokonano mając na uwadze ich zróżnicowane pod względem rozkładu obiektów w przestrzeni pomiarowej, a także liczebność obiektów i parametrów. Nie bez wpływu na wybór danych miał także charakter analizowanego problemu: kalibracja i dyskryminacja. Zastosowano zarówno liniowe i nieliniowe techniki modelowania, dostosowane do analizy problemów o charakterze globalnym oraz lokalnym [55].

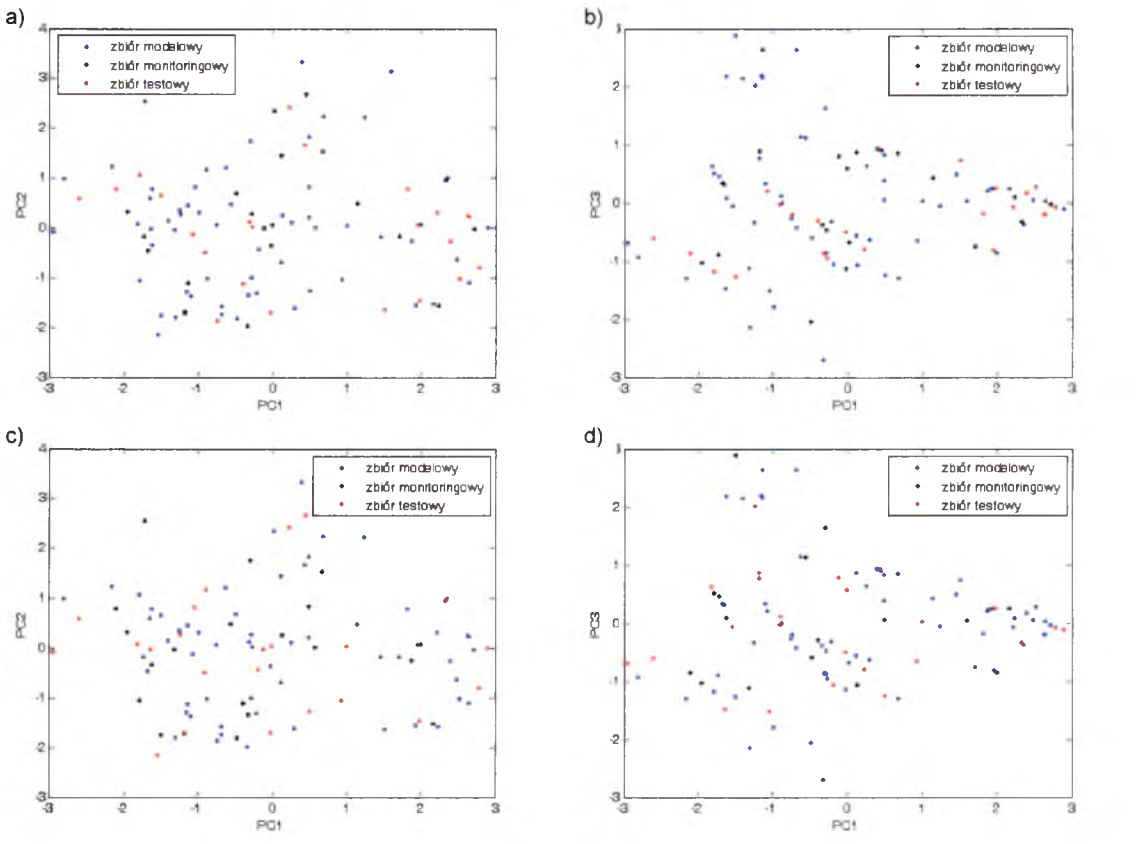
9.1 Dane 1: Modelowanie składu betonu pod względem wytrzymałości

Beton to bardzo skomplikowana mieszanina. Wytrzymałość betonu jest warunkowana nie tylko przez zawartość wody, ale także przez inne czynniki. Dla 103 próbek betonu rejestrowano zawartość 7 składników (w kg na metr sześcienny gotowego produktu), były to: cement, żużel, popiół lotny, woda, domieszka chemiczna (SP), kruszywo grube, kruszywo drobne. Modelowaną własnością była wytrzymałość próbek na ściskanie mierzona w MPa [104].

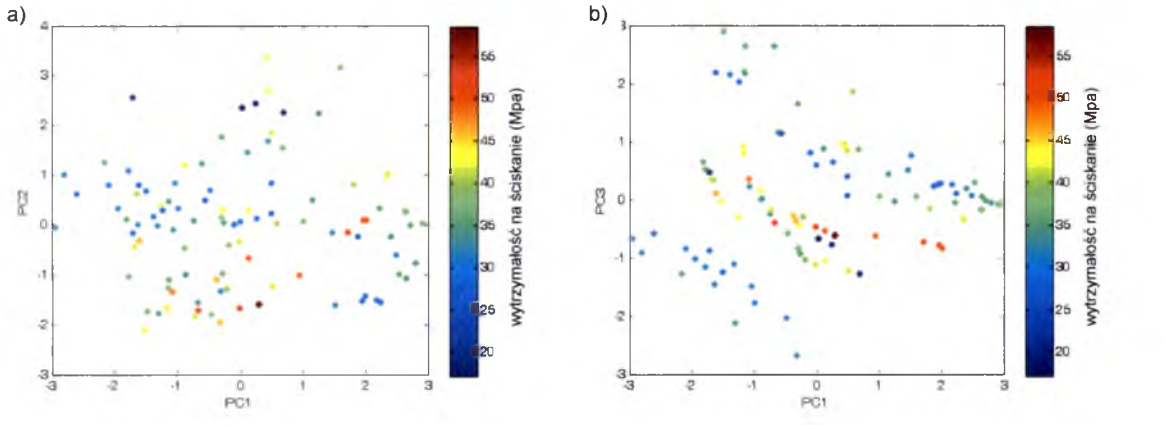
Dane poddane analizie miały wymiarowość 103 x 7. Z uwagi na różny zakres mierzonych parametrów dane zostały poddane autoskalowaniu. Następnie podzielono obiekty na trzy zbiory: 60 próbek do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$), 20 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz 23 do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a oraz algorytmu Duplex. Tak utworzone zbiory ponownie poddano autoskalowaniu wykorzystując wartości średnie oraz odchylenia standardowe dla zbioru modelowego ($\mathbf{X}_{ml}$). Zmienna zależna (wytrzymałość próbek na ściskanie) została poddana centrowaniu, tzn. od każdego elementu wektorów $\mathbf{y}_{ml}$, $\mathbf{y}_{mr}$ i $\mathbf{y}_{tt}$ odjęto wartością średnią ze zbioru modelowego ($\mathbf{y}_{ml}$).

Eksploracja i przygotowanie danych

Celem eksploracji i wizualizacji danych, poddano je analizie czynników głównych (PCA). Kolejne rysunki ukazują rozmieszczenie obiektów należących do zbioru modelowego, monitoringowego i testowego na płaszczyźnie definiowanej przez czynniki główne (Rys. 31a-d). Na projekcji obiektów zdefiniowanej przez pierwsze dwa czynniki główne (Rys. 32a) oraz przez pierwszy i trzeci czynnik główny (Rys. 32b) widoczne jest rozmieszczenie obiektów w zależności od wartości wytrzymałości próbek betonu. Nie można stwierdzić korelacji pomiędzy wartością wytrzymałości próbek betonu na ściskanie, a wartościami czynników głównych. Ponadto w danych nie stwierdzono występowania obiektów odległych. Tak przygotowane dane poddano dalszej analizie – modelowaniu.

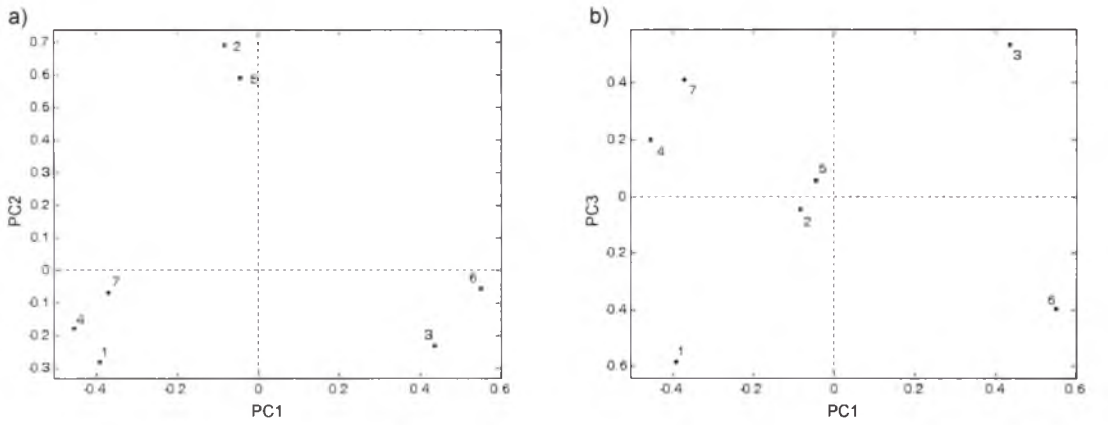


Rys. 31 Projekcja obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



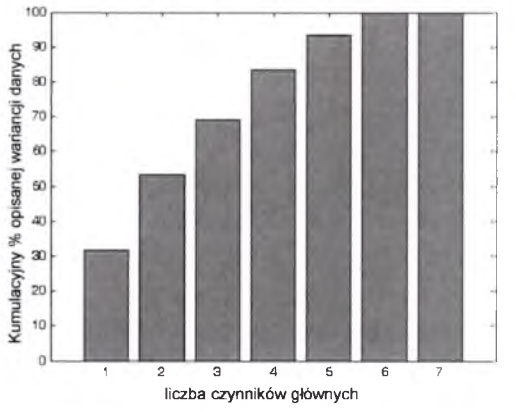
Rys. 32 Projekcja obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy oraz drugi czynnik główny oraz b) przez pierwszy oraz trzeci czynnik główny, gdzie zaznaczono wartości zmiennej zależnej dla każdej próbki

Rys. 33 przedstawia projekcję parametrów na płaszczyznę zdefiniowana przez czynniki główne, gdzie przerywana linia ukazuje początek układu współrzędnych, a o podobieństwie parametrów decyduje kąt pomiędzy wektorami zaczepionymi w początku układu współrzędnych i mającymi koniec w miejscu wyznaczonym przez współrzędne parametru. Analiza projekcji parametrów na płaszczyznę zdefiniowaną przez czynniki główne (Rys. 33) pozwala na stwierdzenie obecności dwóch grup skorelowanych parametrów, są to odpowiednio zmienne 2 i 5 oraz 4 i 7.



Rys. 33 Projekcja parametrów na płaszczyznę zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny, gdzie: 1 – cement, 2 – żużel, 3 – popiół lotny, 4 – woda, 5 – domieszka chemiczna (SP), 6 – kruszywo grube, 7 – kruszywo drobne

Ostatni rysunek przedstawia kumulacyjny procent wariancji danych opisanej przez kolejne czynniki główne (Rys. 34). Widoczne jest, iż kompresja danych nie jest zbyt dobra, co świadczy o niskiej korelacji parametrów zawartych w danych.

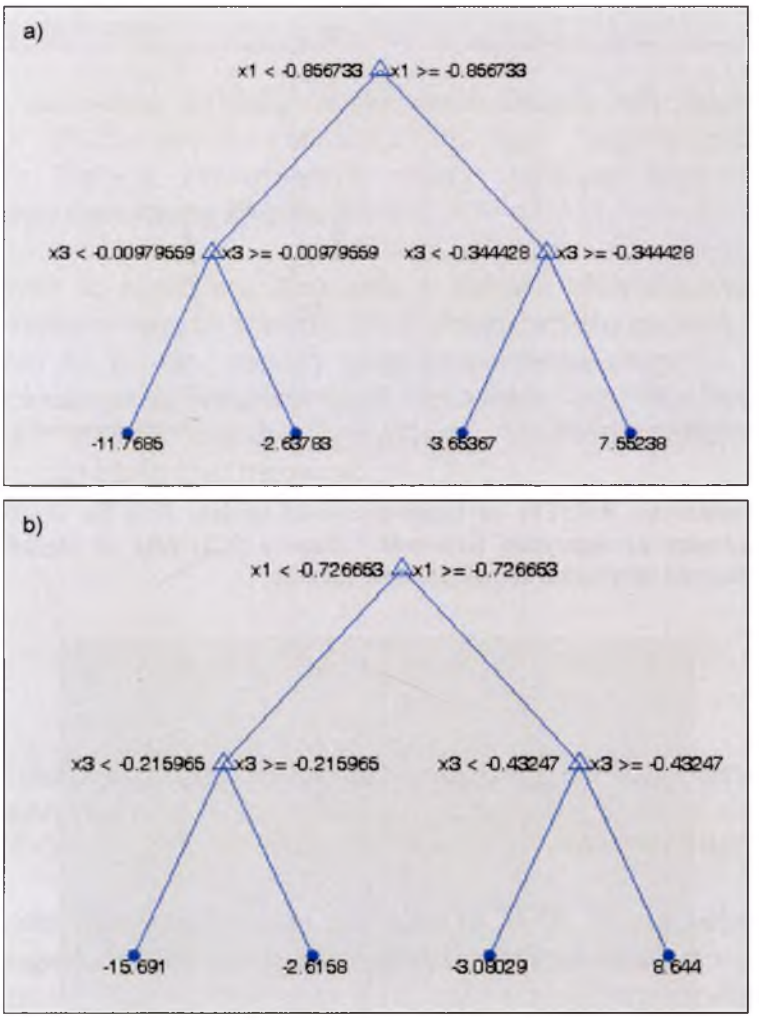


Rys. 34 Kumulacyjny procent wariancji danych opisanej przez kolejne czynniki główne

Poniżej prezentowane są wyniki przeprowadzonych analiz dla danych zawierających zbiory (modelowy, monitoringowy i testowy) skonstruowane za pomocą algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU).

Drzewa klasyfikacji i regresji

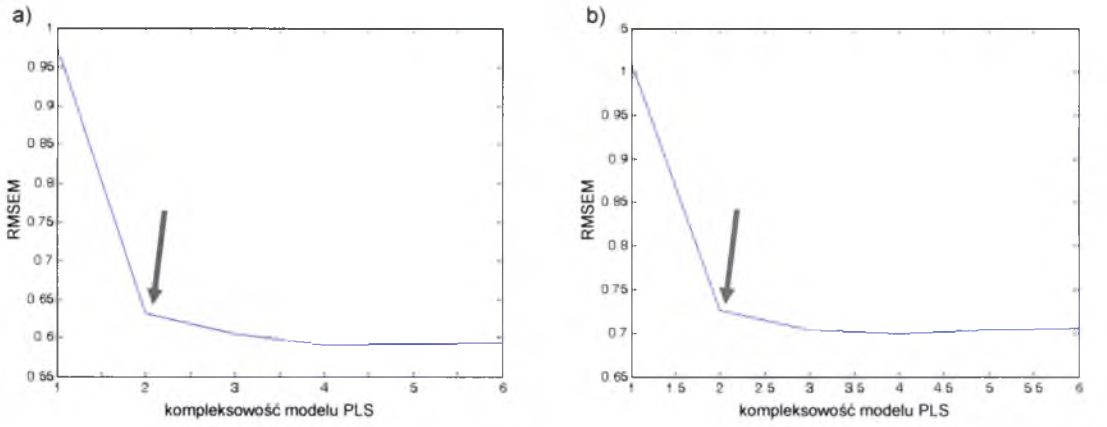
Optymalne binarne drzewa decyzyjne skonstruowane za pomocą metody CART miały cztery węzły terminalne (Rys. 35). Zmienne wybrane przez model CART w oparciu o analizowane dane to zawartość cementu (zmienna 1) oraz popiołu (zmienna 3). Wartości błędów dla optymalnych modeli CART skonstruowanych dla zbiorów danych utworzonych za pomocą zarówno algorytmu Kennarda i Stone'a jak i Duplex wyniosły odpowiednio:
RMSE = 4,40;
RMSEP = 4,00.



Rys. 35 Optymalne drzewo CART skonstruowane celem modelowania wytrzymałości na ściskanie próbek betonu dla zbiorów utworzonych za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU)

Metoda częściowych najmniejszych kwadratów

Druga zastosowana technika modelowania to metoda częściowych najmniejszych kwadratów (PLS). W oparciu o zbiór monitoringowy wybrano po dwa czynniki ukryte do konstrukcji optymalnego modelu (KS, Rys. 36a oraz DU, Rys. 36b). Model ten charakteryzowany był przez następujące wartości pierwiastka średniego błędu kwadratowego:
$RMSE_{(KS)} = 2{,}84$;
$RMSEP_{(KS)} = 2{,}94$
oraz
$RMSE_{(DU)} = 2{,}44$;
$RMSEP_{(DU)} = 2{,}42$.



Rys. 36 Wykres zależności RMSEM od kompleksowości modelu PLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Sieci neuronowe

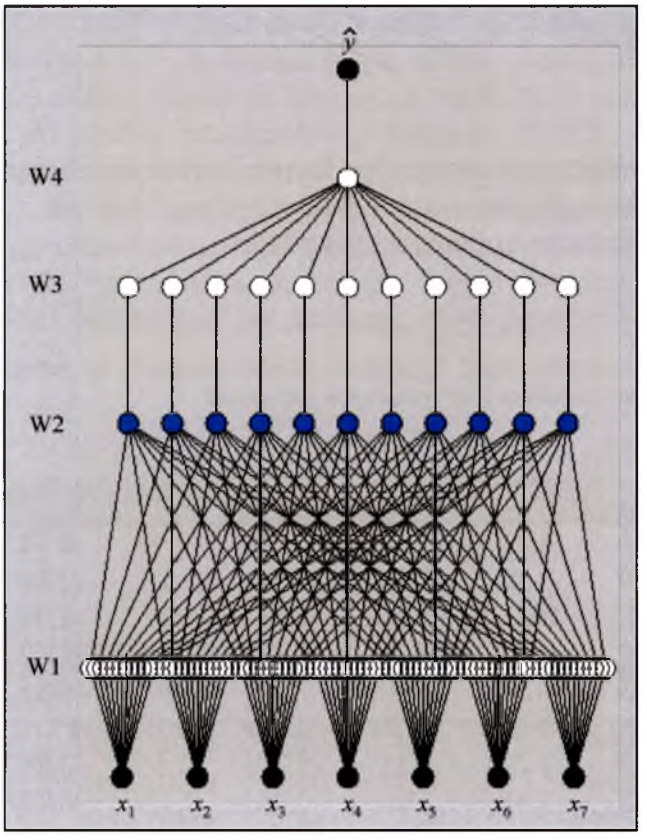
Oryginalne zmienne poddano skalowaniu do przedziału od -1 do 1 co jest niezbędne przed przystąpieniem do modelowania z wykorzystaniem sztucznych sieci neuronowych (ANN) oraz neuronowych systemów rozmytych (NFS).

Konstruując model ANN określa się funkcje aktywacji obecne w węzłach warstwy ukrytej i wyjściowej. W tym przypadku jest to funkcja tangens hiperboliczny w węzłach warstwy ukrytej oraz funkcja liniowa w węźle warstwy wyjściowej. Ponadto optymalizacji wymaga architektura sieci, a więc liczba węzłów w warstwie ukrytej. Najlepsze modele ANN dla modelowania wytrzymałości próbek betonu miały jedenaście węzłów wejściowych oraz jeden węzeł w warstwie wyjściowej. Ilość węzłów w warstwie ukrytej sieci może być różna w zależności od sposobu przypisywania próbek do odpowiednich zbiorów. Sieć konstruowana celem modelowania danych zawierających zbiory z algorytmu Kennarda i Stone'a miała trzy węzły w warstwie ukrytej. Warstwa ukryta sieci dla modelowania danych zawierających zbiory z algorytmu Duplex miała cztery węzły w warstwie ukrytej. Optymalny model sztucznych sieci neuronowych charakteryzowany był przez następujące wartości pierwiastka średniego błędu kwadratowego:
$RMSE_{(KS)} = 0{,}28$;
$RMSEP_{(KS)} = 0{,}44$
oraz
$RMSE_{(DU)} = 0{,}34$;
$RMSEP_{(DU)} = 0{,}54$.

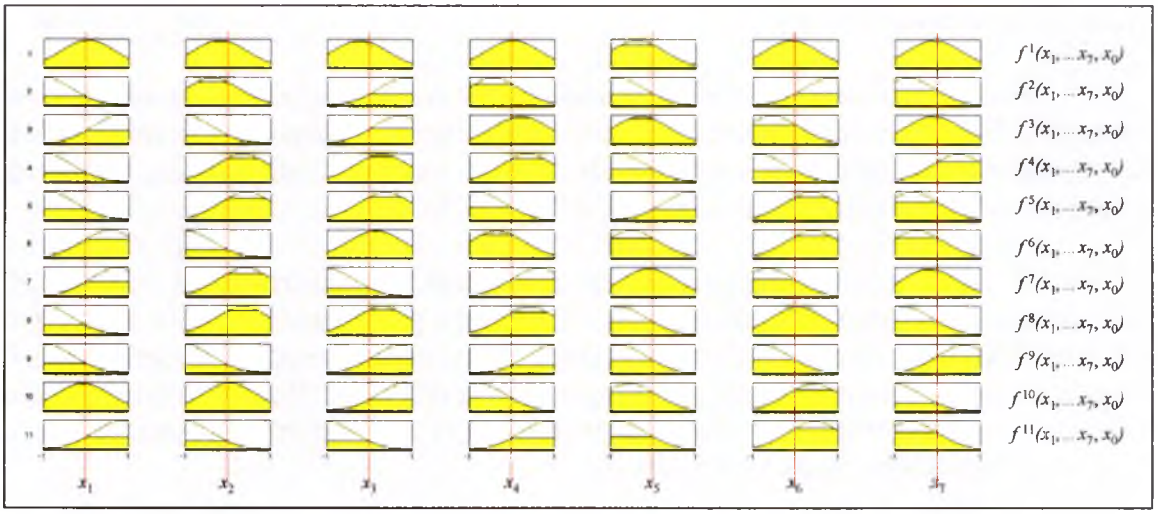
Neuronowe systemy rozmyte

Ostatnią stosowaną techniką modelowania danych były neuronowe systemy rozmyte (NFS). Skonstruowano modele NFS typu Sugeno pierwszego rzędu do modelowania danych zawierających zbiory uzyskane algorytmem Kennarda i Stone'a (KS) oraz algorytmem Duplex (DU).

Dla danych z algorytmu Kennarda i Stone'a skonstruowano model NFS o strukturze zaprezentowanej na rysunku 37. Podziału przestrzeni danych w niniejszym modelu dokonano za pomocą metody grupowania różnicowego (o promieniu 0,9), pozwoliło to na konstrukcję jedenastu reguł logicznych (Rys. 38). Neuronowy system rozmyty uczono według metody hybrydowej. Uzyskano następujące wartości pierwiastka średniego błędu kwadratowego:
$RMSE_{(KS)} = 0{,}00$;
$RMSEP_{(KS)} = 1{,}18$.



Rys. 37 Struktura modelu NFS dla modelowania składu betonu pod względem wytrzymałości, gdzie x_i oznacza *i*-ty parametr: 1 – cement, 2 – żużel, 3 – popiół lotny, 4 – woda, 5 – domieszka chemiczna (SP), 6 – kruszywo grube, 7 – kruszywo drobne, natomiast W*j* oznacza *j*-tą warstwę opisaną w rozdziale *7 Neuronowe systemy rozmyte*.



Rys. 38 Jedenaście reguł logicznych skonstruowanych w ramach modelu NFS celem modelowania składu betonu pod względem jego wytrzymałości, gdzie w poziomie znajdują się reguły logiczne, w pionie funkcje przynależności przypadające na każdy parametr (x_i), a ostatnia kolumna to liniowe kombinacje oryginalnych zmiennych

Dla tego konkretnego przypadku linowe kombinacje oryginalnych zmiennych to jednomian siódmego stopnia mający postać ogólną $f^l(\mathbf{x}) = \mathbf{b}\ \mathbf{x} + \mathbf{b}_0$, gdzie wektory współczynników **b** zamieszczono poniższej tabeli.

Tabela 1 Następniki jedenastu reguł logicznych skonstruowanych w ramach modelu NFS będące współczynnikami liniowej kombinacji oryginalnych zmiennych

	współczynniki dla kolejnych zmiennych i wyraz wolny							
reguła 1	6.70	4.86	12.90	-11.84	-0.98	-4.74	-6.38	1.80
reguła 2	8.70	-1.42	5.86	-3.51	2.15	-2.80	2.49	2.40
reguła 3	13.43	10.02	12□62	2.36	5.37	-1.94	-2.07	0.06
reguła 4	4.10	-3.32	10.52	2.71	-7.90	-3.62	1.81	-14.10
reguła 5	8.63	-1.83	15.45	-7.42	3.59	-6.55	-3.40	-3.80
reguła 6	29.96	-9.21	11.72	-10.53	2.61	-17.62	1.95	-12.17
reguła 7	5.40	7.22	3.35	-5.22	-0.65	-3.39	-2.49	-7.22
reguła 8	0.53	8.48	1.96	-0.26	7.12	-6.13	0.09	-7.56
reguła 9	-3.45	-2.08	1.63	4.62	5.27	6.19	-6.00	1.48
reguła 10	10.01	-3.22	6.02	-17.63	5.59	-24.50	22.52	35.05
reguła 11	-16.82	-0.54	11.23	-0.38	-7.05	-12.04	4.87	26.27

Następnie skonstruowano model NFS dla danych zawierających zbiory utworzone za pomocą algorytmu Duplex. Otrzymany model NFS dostarczył reguł logicznych składających się z funkcji przynależności (część poprzednika) oraz liniowych kombinacji oryginalnych zmiennych (część następnika). Jednakże z uwagi na dużą obszerność i mały wkład tej informacji w analizę wyników rysunki przedstawiające reguły logiczne (jak Rys. 38), strukturę NFS (jak Rys. 37) oraz tabela

z następnikami reguł logicznych (jak Tabela 1) nie będą prezentowane w toku dalszej analizy kolejnych zestawów danych. Podane zostaną jedynie istotne parametry konstruowanego modelu, jak typ podziału przestrzeni pomiarowej i metoda uczenia modelu.

Dla danych z algorytmu Duplex podział przestrzeni danych w modelu NFS został przeprowadzony według metody grupowania różnicowego (o promieniu 0,6), która pozwoliła na konstrukcję dwudziestu sześciu reguł logicznych (Rys. 24a). W tym przypadku model NFS był także uczony według metody hybrydowej. Uzyskano następujące wartości pierwiastka średniego błędu kwadratowego:
$RMSE_{(DU)} = 0{,}00$;
$RMSEP_{(DU)} = 0{,}89$.

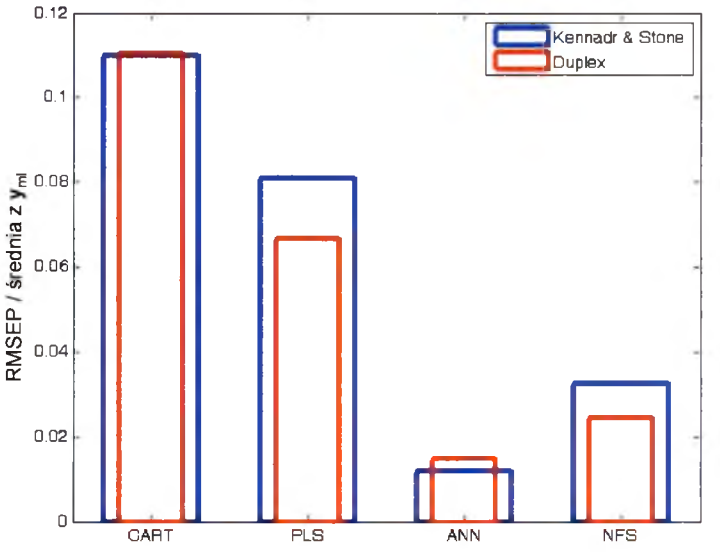
Podsumowanie

W poniższej tabeli zestawiono wyniki modelowania składu betonu pod względem wytrzymałości na ściskanie (w MPa). Zastosowano cztery metody modelowania danych, mające charakter liniowy (CART, PLS) oraz nieliniowy (ANN, NFS), dostosowane do analizy problemów globalnych (CART, PLS) lub lokalnych (NFS). Wszystkie modele konstruowano w oparciu o oryginalne zmienne. Kolumna czwarta w tabeli 2 zawiera wartości pierwiastka średniego błędu kwadratowego (RMSE) obrazującego odpasowanie modelu do danych, natomiast w kolumnie piątej zamieszczono wartości pierwiastka średniego błędu kwadratowego dla próbek z niezależnego zbioru testowego, co obrazuje moc predykcyjną konstruowanych modeli.

Tabela 2 Zestawienie wyników przeprowadzonych analiz dla modelowani składu betonu względem wytrzymałości (Dane 1), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	RMSE	RMSEP	opis modelu
CART	KS	oryginalne	4,40	4,00	4 węzły terminalne
	DU	oryginalne	4,40	4,00	4 węzły terminalne
PLS	KS	oryginalne	2,84	2,94	2 czynniki ukryte
	DU	oryginalne	2,44	2,42	2 czynniki ukryte
ANN	KS	oryginalne	0,28	0,44	3 węzły w warstwie ukrytej
	DU	oryginalne	0,34	0,54	4 węzły w warstwie ukrytej
NFS	KS	oryginalne	0,00	1,18	11 reguł logicznych
	DU	oryginalne	0,00	0,89	26 reguł logicznych

Podsumowanie otrzymanych wyników przedstawia rysunek 39. Wartości błędu zostały podzielone przez wartość średnią zmiennej zależnej ze zbioru modelowego celem porównania wyników dla różnych zestawów danych.

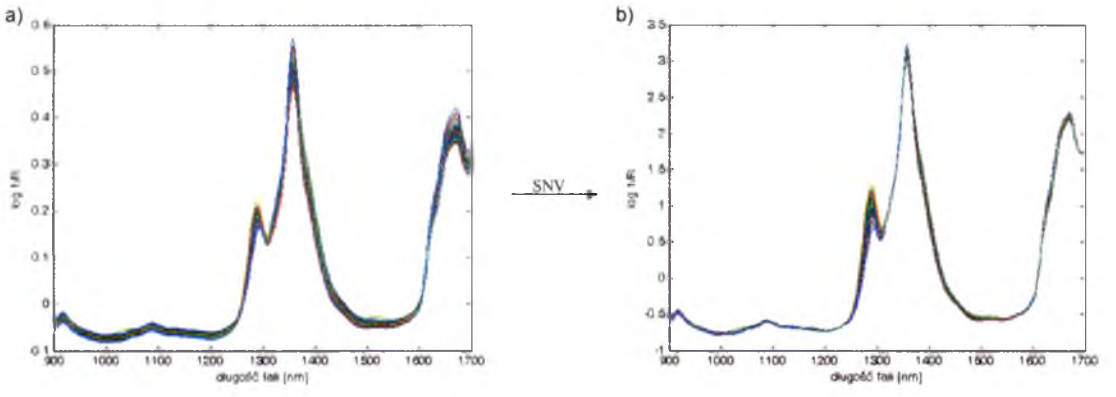


Rys. 39 Wykres wartości pierwiastka średniego błędu kwadratowego charakteryzujący konstruowane modele celem modelowania wytrzymałości na ściskanie próbek betonu

Porównując otrzymane wyniki dla modelu NFS z wynikami dla modelu CART można zauważyć, iż neuronowe systemy rozmyte dały model obarczony niższym błędem przewidywania dla próbek z niezależnego zbioru testowego, jednocześnie dostarczając reguły logiczne. Model NFS odznacza się także mniejszym błędem w porównaniu do powszechnie stosowanej w chemii metody modelowania danych to jest metody częściowych najmniejszych kwadratów (PLS). Metoda PLS nie dostarcza jednak reguł logicznych. Sieci neuronowe pozwoliły na konstrukcję nieco lepszego modelu niż model NFS.

9.2 Dane 2: Modelowanie liczby oktanowej próbek benzyny

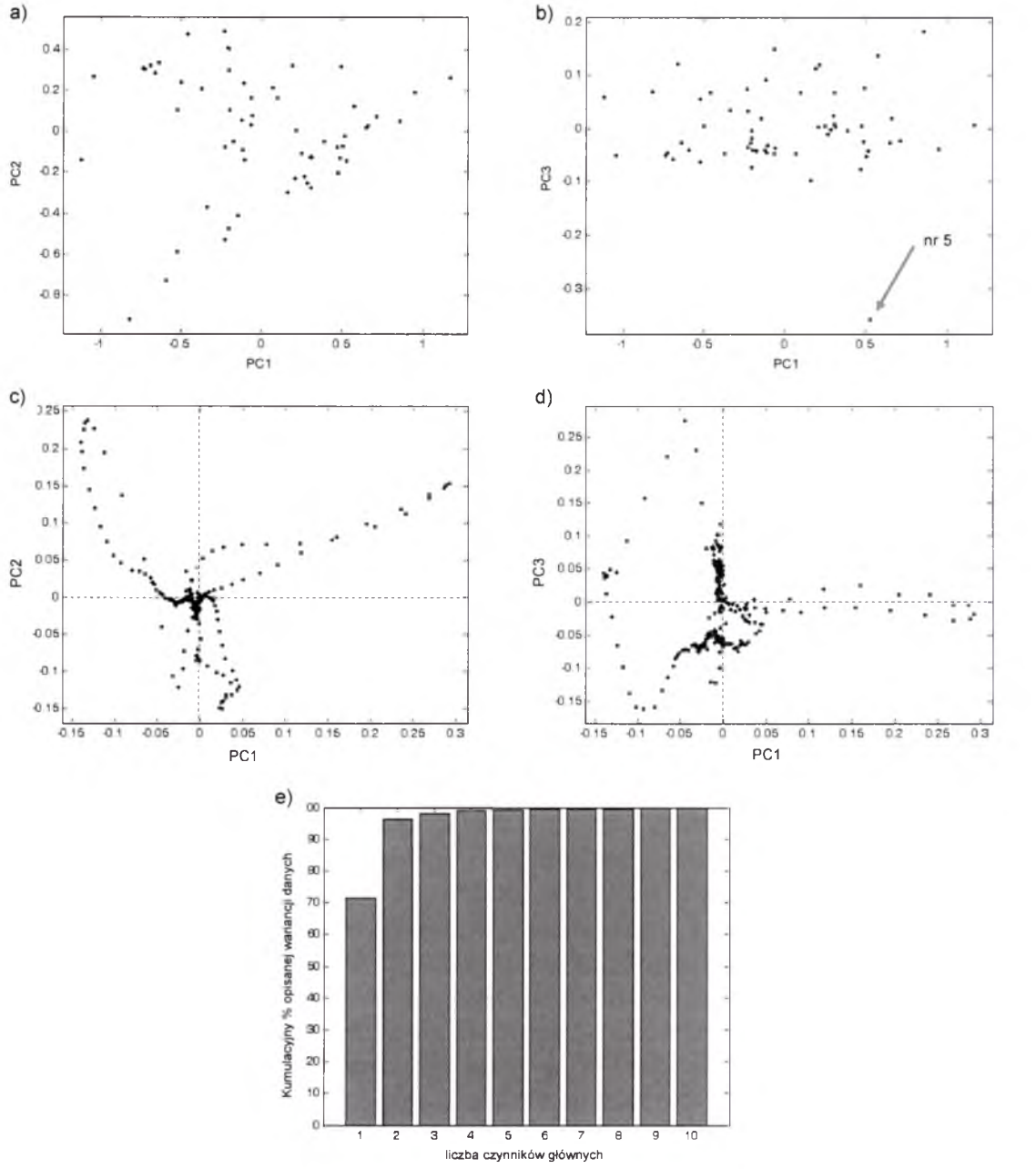
Dane 2 zawierały widma w bliskiej podczerwieni dla 60 próbek benzyny [105]. Widma NIR rejestrowano w wariancie odbiciowym (log 1/R) w zakresie od 900 nm do 1700 nm. Dla każdej z próbek oznaczono metodą referencyjną liczbę oktanową. Dane o wymiarowości 60 x 256 poddano wstępnej obróbce stosując transformację SNV (Rys. 40).



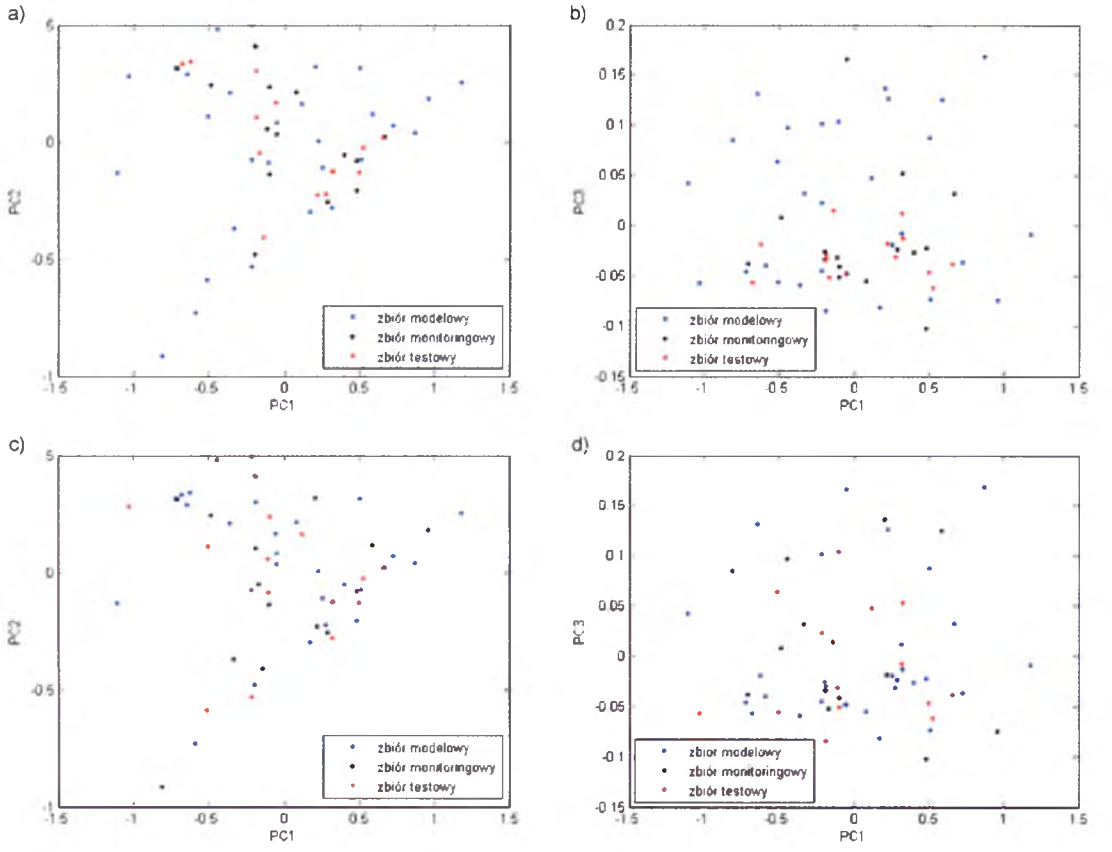
Rys. 40 Widma NIR 60 próbek benzyny a) przed i b) po transformacji SNV

Eksploracja i przygotowanie danych

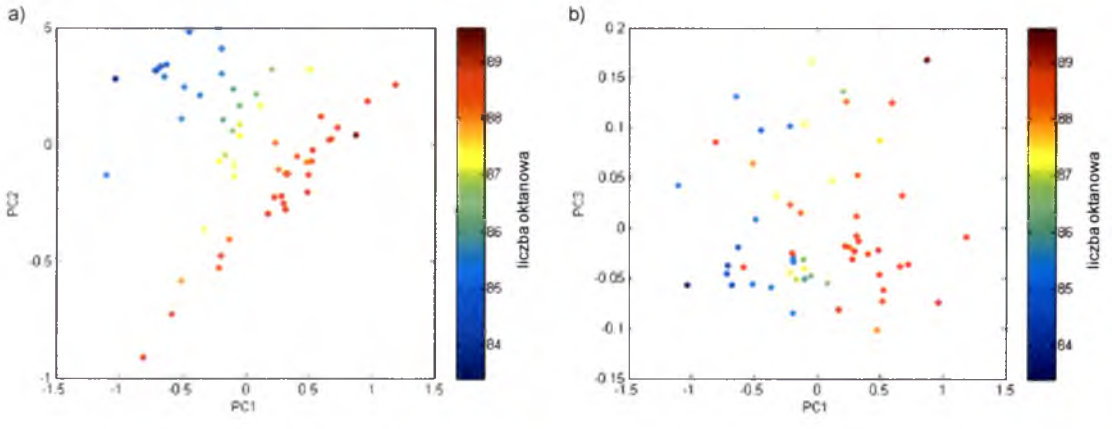
Zastosowano analizę czynników głównych celem wizualizacji i eksploracji danych. W toku analizy wykryto jeden obiekt odległy, była to próbka nr 5 (Rys. 41b). Obiekty odległe nie są pożądanymi elementami składowymi danych z uwagi na fakt, iż mogą negatywnie wpływać na właściwości predykcyjne konstruowanych modeli. Próbka nr 5 została usunięta. W wyniku ponownie przeprowadzonej analizy PCA (Rys. 42, Rys. 44) nie stwierdzono obecności innych obiektów odległych.



Rys. 41 Projekcja 60 obiektów na płaszczyznę zdefiniowaną przez a) pierwszy i drugi czynnik główny, oraz b) pierwszy i trzeci czynnik główny (strzałka wskazuje obiekt odległy); projekcja parametrów na płaszczyznę zdefiniowaną przez c) pierwszy i drugi czynnik główny, oraz d) pierwszy i trzeci czynnik główny; e) kumulacyjny procent wariancji danych opisanej przez kolejne czynniki główne



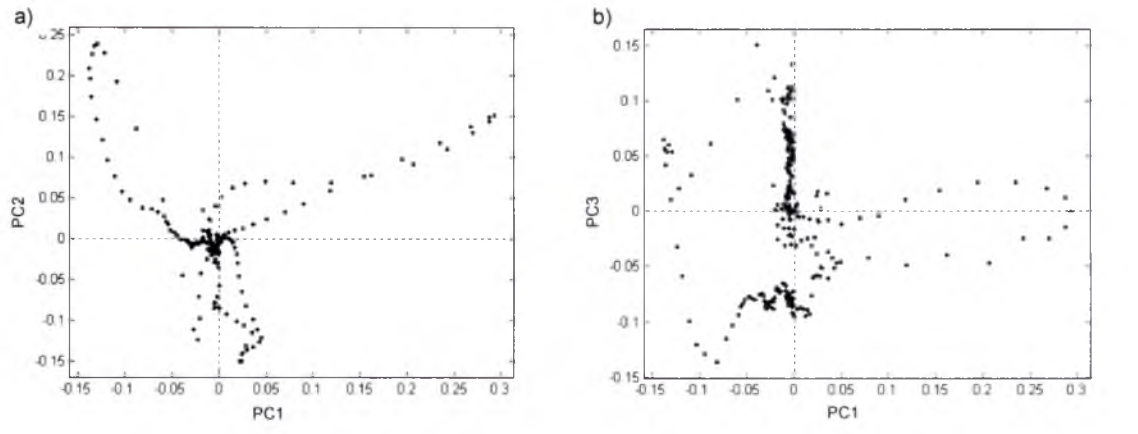
Rys. 42 Projekcja 59 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 43 Projekcja obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono liczbę oktanową dla każdej próbki

Projekcja obiektów na płaszczyzny zdefiniowane przez czynniki główne (Rys. 43) ujawniła zależność pomiędzy liczbą oktanową próbki benzyny, a wartościami czynników głównych. Szczególnie na rysunku 43a widoczne jest iż liczba oktanowa jest wprostproporcionalna do wartości czynnika pierwszego oraz odwrotnie proporcjonalna do wartości drugiego czynnika.

Eliminacja obiektu odległego nr 5 wpłynęła także nieznacznie na rozmieszczenie parametrów w przestrzeni zdefiniowanej przez czynniki główne (Rys. 44).

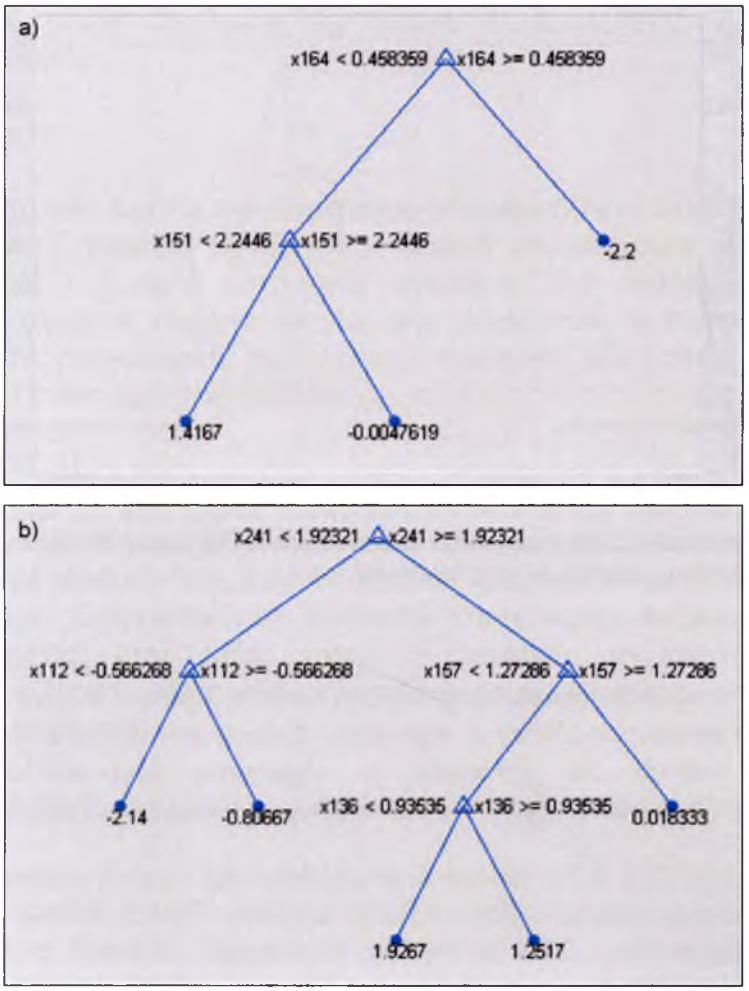


Rys. 44 Projekcja parametrów na płaszczyznę wag zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny

Następnie 59 obiektów podzielono na trzy zbiory przypisując 30 obiektów do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$), 15 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz 14 do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Zmienna zależna dla wszystkich zbiorów ($\mathbf{y}_{ml}$, $\mathbf{y}_{mr}$, $\mathbf{y}_{tt}$) została poddana centrowaniu. Tak utworzone zbiory zostały poddane modelowaniu metodą CART oraz PLS.

Drzewa klasyfikacji i regresji

Następnie przystąpiono do konstrukcji modelu drzew klasyfikacji i regresji (CART). Optymalne binarne drzewo decyzyjne konstruowane w oparciu o zbiory tworzone za pomocą algorytmu Kennarda i Stone'a miało trzy węzły terminalne (Rys. 45a). Zmienne wskazane w modelu jako decyzyjne to zmienna 151 i 164 oraz zmienna 137 wskazana przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły odpowiednio:
$RMSE_{(KS)} = 0{,}45$;
$RMSEP_{(KS)} = 0{,}46$.

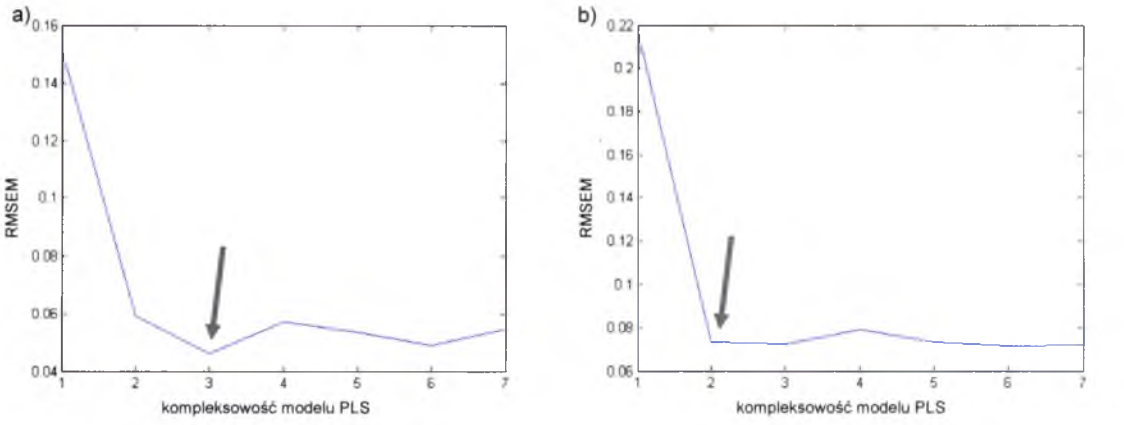


Rys. 45 Optymalne drzewo CART skonstruowane celem modelowania liczby oktanowej próbek benzyny dla zbiorów utworzonych za pomocą a) algorytmu Kennarda i Stone'a oraz b) algorytmu Duplex

Natomiast optymalny model konstruowany w oparciu o dane zawierające zbiory utworzone za pomocą algorytmu Duplex miał pięć węzłów terminalnych (Rys. 45b). Zmienne wskazane w modelu jako decyzyjne to zmienna 112, 136, 157, 241 oraz zmienna 151 wskazana przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły odpowiednio:
$RMSE_{(DU)} = 0{,}28$;
$RMSEP_{(DU)} = 0{,}63$.

Metoda częściowych najmniejszych kwadratów

Druga wykorzystana technika modelowania danych to metoda częściowych najmniejszych kwadratów. Wyznaczono kompleksowość modelu PLS – wybrano trzy czynniki ukryte do konstrukcji optymalnego modelu konstruowanego w oparciu o dane zawierające zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 46a) oraz dwa czynniki ukryte dla modelowania danych zawierających zbiory otrzymane za pomocą algorytmu Duplex (DU, Rys. 46b).



Rys. 46 Wykres zależności RMSEM od kompleksowości modelu PLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Finalny model skonstruowany odpowiednio w oparciu o dwa czynniki główne dla danych zawierających zbiory z algorytmu Kennarda i Stone'a oraz trzy czynniki główne dla danych ze zbiorami z algorytmu Duplex charakteryzowany był przez następujące wartości pierwiastka średniego błędu kwadratowego:
$RMSE_{(KS)} = 0,21$;
$RMSEP_{(KS)} = 0,17$
oraz
$RMSE_{(DU)} = 0,16$;
$RMSEP_{(DU)} = 0,30$.

Sieci neuronowe

Konstrukcja modelu w ramach metody CART oraz PLS została przeprowadzona w oparciu o całe widma NIR. Dla tych samych danych konstrukcja modelu ANN i NFS opierała się o skompresowane dane. Redukcję wymiarowości danych można przeprowadzić przez zastąpienie oryginalnych zmiennych czynnikami głównymi, albo zmiennymi istotnymi. Czynniki główne otrzymuje się stosując metodę PCA, a zmienne istotne pochodzą z algorytmów wyboru zmiennych istotnych. Istnieją różne metody wyboru zmiennych istotnych [106]. W niniejszej pracy do konstrukcji modeli ANN oraz NFS wykorzystywane są czynniki główne, a także zmienne wybrane przez model CART. Dodatkowo nowe zredukowane zmienne poddano skalowaniu do przedziału <-1, 1>.

Konstruowany model ANN zawierał we wszystkich węzłach warstwy ukrytej funkcję typu tangens hiperboliczny, a w węźle warstwy wyjściowej funkcję liniową. Jako pierwszy modelowany zestaw danych użyto czterech czynników głównych (PCs) opisujących 99,17% wariancji danych podzielonych na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalna sieć zawierała cztery węzły wejściowe, trzy węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej. Sieć ta pozwoliła

na przewidzenie liczby oktanowej dla próbek z niezależnego zbioru testowego z następującymi błędami:
$RMSE_{(KS/4PCs)} = 0,25$;
$RMSEP_{(KS/4PCs)} = 0,16$.

Kolejny zestaw danych zawierał zmienne istotne (ZM: 137, 151, 164) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Optymalny model to sieć zawierająca trzy węzły wejściowe, trzy węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej. Pozwoliła ona na przewidzenie modelowanej własności dla próbek z niezależnego zbioru testowego z następującymi błędami:
$RMSE_{(KS/3ZM)} = 0,52$;
$RMSEP_{(KS/3ZM)} = 0,21$.

Następny modelowany zestaw danych to cztery czynniki główne (PCs) opisujące 99,17% wariancji danych, dla których obiekty podzielono na zbiory za pomocą algorytmu Duplex. Optymalna sieć zawierała cztery węzły wejściowe, trzy węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej i pozwoliła na przewidzenie liczby oktanowej dla próbek z niezależnego zbioru testowego z następującymi błędami:
$RMSE_{(DU/4PCs)} = 0,23$;
$RMSEP_{(DU/4PCs)} = 0,18$.

Ostatni zestaw danych zawierał zmienne istotne (ZM: 112, 136, 151, 157, 241) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Optymalny model to sieć zawierająca pięć węzłów wejściowych, trzy węzły w warstwie ukrytej i jeden węzeł w warstwie wyjściowej. Sieć ta pozwoliła na przewidzenie modelowanej własności dla próbek z niezależnego zbioru testowego z następującymi błędami:
$RMSE_{(DU/5ZM)} = 0,23$;
$RMSEP_{(DU/5ZM)} = 0,23$.

Neuronowe systemy rozmyte

Jako ostatnią technikę modelowania danych zastosowano neuronowe systemy rozmyte. Skonstruowano modele NFS typu Sugeno pierwszego rzędu do modelowania danych zawierających zbiory uzyskane algorytmem Kennarda i Stone'a (KS) oraz algorytmem Duplex (DU). Jako pierwszy modelowany zestaw danych zostały użyte cztery czynniki główne opisujące 99,17% wariancji danych, dla których obiekty podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Iteracyjne uczenie modelu odbywało się w oparciu o model hybrydowy. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Skonstruowany model pozwolił na przewidzenie liczby oktanowej z następującymi błędami:
$RMSE_{(KS/4PCs)} = 0,21$;
$RMSEP_{(KS/4PCs)} = 0,23$.

Drugi zestaw danych zawierał zmienne istotne (ZM: 137, 151, 164) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model wykorzystywał metodę kratkową do podziału przestrzeni danych. W ramach tego modelu skonstruowano 64 reguły logiczne poprzez przypisanie każdej zmiennej czterech funkcji przynależności. Konstruowany model obarczony był błędami:
$RMSE_{(KS/3ZM)} = 0{,}00$;
$RMSEP_{(KS/3ZM)} = 0{,}25$.

Następny modelowany zestaw danych to cztery czynniki główne (PCs) opisujące 99,17% wariancji danych. Obiekty podzielono na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano pięć reguł logicznych. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Skonstruowany model pozwolił na przewidzenie liczby oktanowej z następującymi błędami:
$RMSE_{(DU/4PCs)} = 0{,}03$;
$RMSEP_{(DU/4PCs)} = 0{,}26$.

Czwarty zestaw danych zawierał zmienne istotne (ZM: 112, 136, 151, 157, 241) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Iteracyjne uczenie modelu odbywało się w oparciu o wsteczną propagację błędu. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowanych było sześć reguł logicznych. Konstruowany model obarczony był następującymi błędami:
$RMSE_{(DU/5ZM)} = 0{,}00$;
$RMSEP_{(DU/5ZM)} = 0{,}25$.

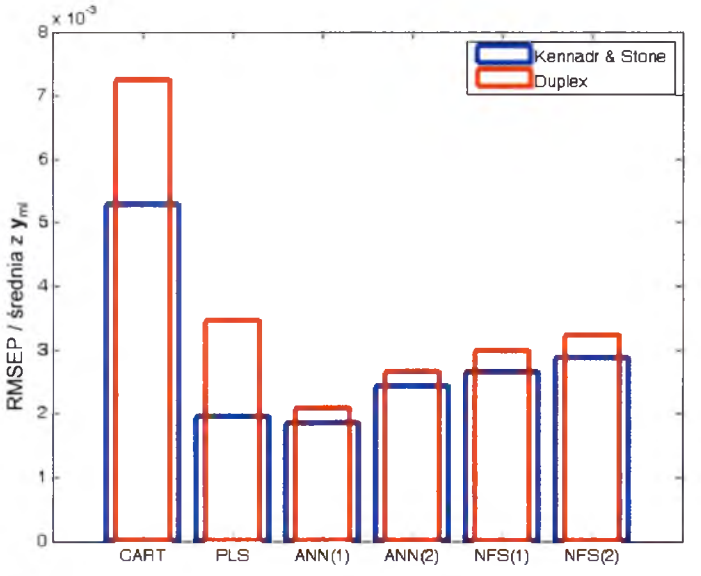
Podsumowanie

Tabela 3 zawiera wyniki modelowania liczby oktanowej próbek benzyny za pomocą czterech różnych technik chemometrycznych. Modele CART i PLS skonstruowano w oparciu o oryginalne zmienne. Natomiast do konstrukcji modeli ANN i NFS wykorzystano dane skompresowane: czynniki główne (PCs) oraz wybrane zmienne istotne (ZM). W tabeli zamieszczono wartości błędów RMSE oraz RMSEP obrazujące odpowiednio odpasowanie modelu do danych i moc predykcyjną konstruowanych modeli.

Tabela 3 Zestawienie wyników przeprowadzonych analiz dla modelowani liczby oktanowej benzyny (Dane 2), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	RMSE	RMSEP	opis modelu
CART	KS	oryginalne	0,45	0,46	3 węzły terminalne
	DU	oryginalne	0,28	0,63	5 węzłów terminalnych
PLS	KS	oryginalne	0,21	0,17	3 czynniki ukryte
	DU	oryginalne	0,16	0,30	2 czynniki ukryte
ANN	KS	4 PCs	0,25	0,16	3 węzły w warstwie ukrytej
		3 ZM	0,52	0,21	3 węzły w warstwie ukrytej
	DU	4 PCs	0,23	0,18	3 węzły w warstwie ukrytej
		5 ZM	0,23	0,23	3 węzły w warstwie ukrytej
NFS	KS	4 PCs	0,21	0,23	2 reguły logiczne
		3 ZM	0,00	0,25	64 reguły logiczne
	DU	4 PCs	0,03	0,26	5 reguł logicznych
		5 ZM	0,00	0,25	6 reguł logicznych

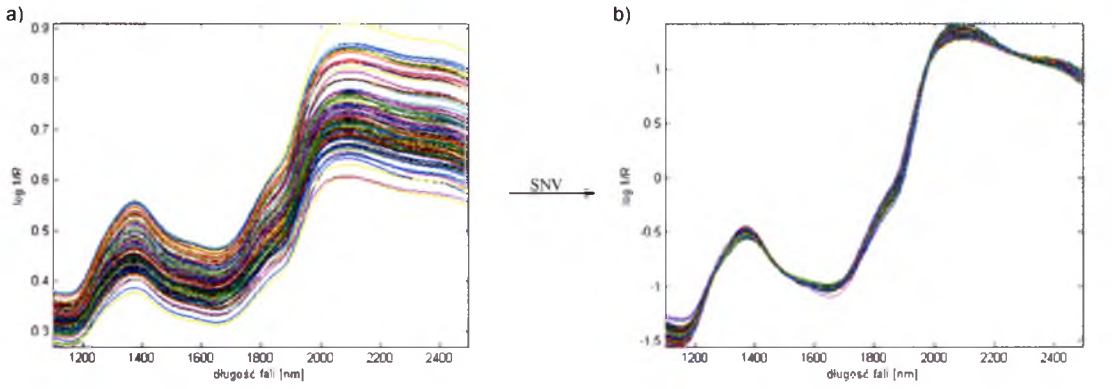
Na Rys. 47 przedstawiono wykres zawierający otrzymane wyniki modelowania za pomocą zastosowanych metod.



Rys. 47 Wykres wartości pierwiastka średniego błędu kwadratowego charakteryzujący konstruowane modele celem modelowania liczby oktanowej próbek benzyny, gdzie indeksy oznaczają modele konstruowane w oparciu odpowiednio o dane zawierające (1) czynniki główne oraz (2) zmienne istotne

Porównując otrzymane wyniki dla modelu NFS i modelu CART, można zauważyć, iż neuronowe systemy rozmyte dały model obarczony niższym błędem jednocześnie dostarczając reguły logiczne. Model NFS cechował się także minimalnie gorszym błędem od błędu powszechnie stosowanych w chemii metod modelowania danych PLS oraz ANN, które nie dostarczają reguł logicznych.

9.3 Dane 3: Modelowanie zawartości wilgoci w pszenicy

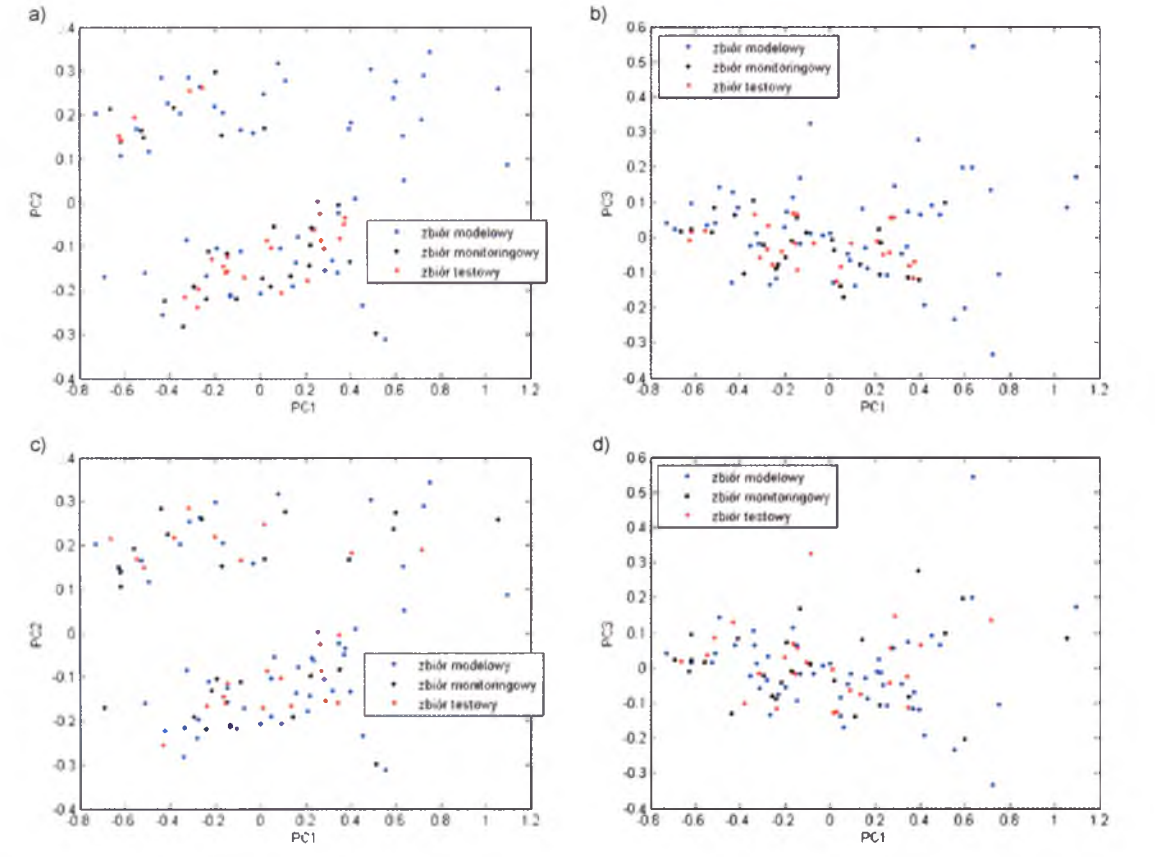


Rys. 48 Widma NIR 100 próbek pszenicy a) przed i b) po transformacji SNV

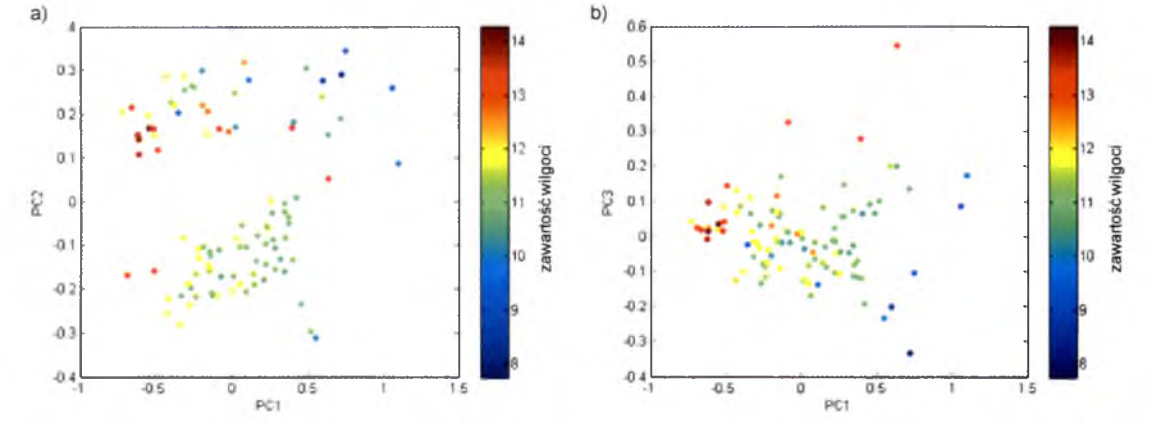
Dane 3 zawierały widma zarejestrowane techniką spektroskopii w bliskiej podczerwieni dla 100 próbek pszenicy [105]. Widma rejestrowano spektroskopowo w wariancie odbiciowym (log 1/R) w zakresie od 1100 nm do 2500 nm. Dla każdej z próbek metodą referencyjną oznaczono zawartość wilgoci. Dane o wymiarowości 100 x 256 poddano wstępnej obróbce stosując transformacji SNV (Rys. 48).

Eksploracja i przygotowanie danych

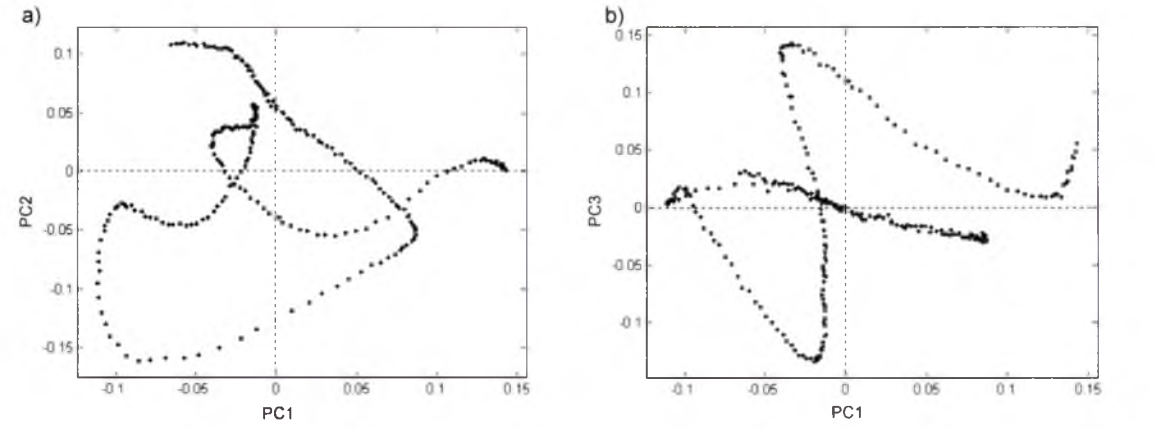
Analiza PCA nie ujawniła żadnych nieprawidłowości w danych, w tym obiektów odległych (Rys. 49, 51, 52). Nie stwierdzono także zależności pomiędzy modelowaną własnością a rozmieszczeniem obiektów w przestrzeni czynników głównych (Rys. 50).



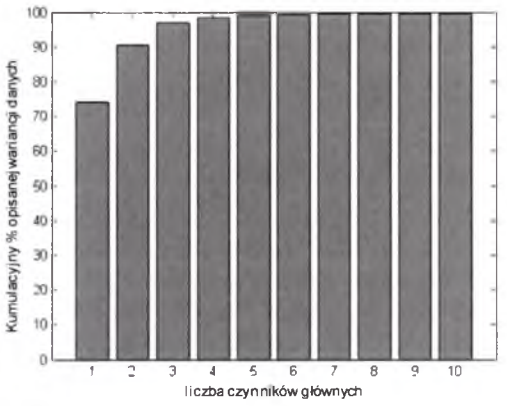
Rys. 49 Projekcja 100 próbek pszenicy na płaszczyznę zdefiniowaną przez a, c) pierwszy i drugi czynnik główny oraz b, d) pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 50 Projekcja 100 próbek pszenicy na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono zawartość wilgoci w każdej próbce



Rys. 51 Projekcja parametrów na płaszczyznę zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny



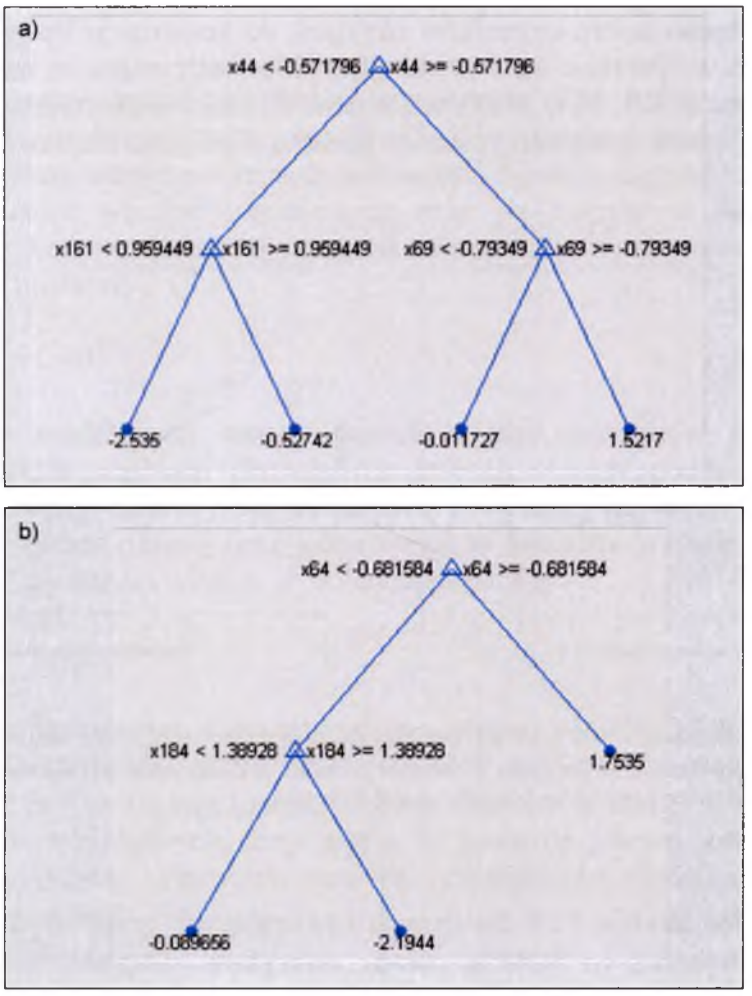
Rys. 52 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

Następnie 100 próbek pszenicy podzielono na trzy zbiory przypisując 50 obiektów do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$) oraz po 25 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) i testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Zmienna zależna dla wszystkich zbiorów ($\mathbf{y}_{ml}$, $\mathbf{y}_{mr}$, $\mathbf{y}_{tt}$) została poddana centrowaniu. Tak utworzone zbiory zostały poddane modelowaniu metodą CART oraz PLS.

Drzewa klasyfikacji i regresji

Jako pierwszą metodę modelowania danych zastosowano metodę CART. Optymalne binarne drzewo decyzyjne konstruowane w oparciu o zbiory utworzone za pomocą algorytmu Kennarda i Stone'a miało cztery węzły terminalne (Rys. 53a). Zmienne wskazane w modelu jako decyzyjne to zmienne 44, 161 i 169 oraz zmienne 1, 36, 76, 81, 96, 99, 207 wskazane przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły odpowiednio:

$RMSE_{(KS)} = 0{,}64$;
$RMSEP_{(KS)} = 0{,}66$.

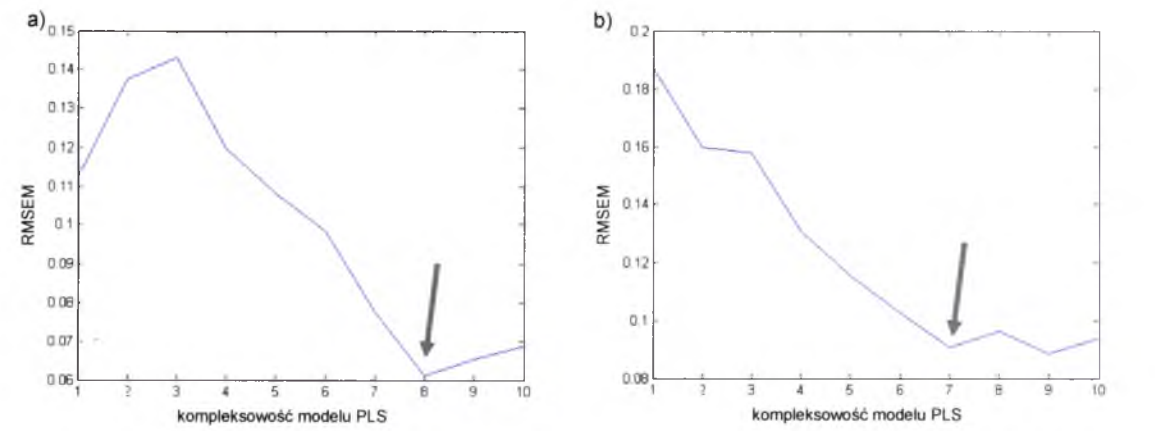


Rys. 53 Optymalne drzewo CART skonstruowane celem modelowania zawartości wilgoci w pszenicy dla zbiorów utworzonych za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU)

Dla danych zawierających zbiory utworzone za pomocą algorytmu Duplex optymalny model miał trzy węzły terminalne (Rys. 53b). Zmienne wskazane w modelu jako decyzyjne to zmienne 64 i 184 oraz zmienne 29, 37, 62, 162, 167, 224, 239 wskazane przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły odpowiednio:
$RMSE_{(DU)} = 0{,}65$;
$RMSEP_{(DU)} = 0{,}70$.

Metoda częściowych najmniejszych kwadratów

Dla analizowanych danych skonstruowano model PLS, dla którego wyznaczono kompleksowość w oparciu o błąd przewidywania dla zbioru monitoringowego (RYSEM). Wybrano osiem czynników ukrytych do konstrukcji optymalnego modelu konstruowanego w oparciu o dane zawierające zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 54a) oraz siedem czynników ukrytych dla modelowania danych zawierających zbiory otrzymane za pomocą algorytmu Duplex (DU, Rys. 54b).



Rys. 54 Wykres zależności RMSEM od kompleksowości modelu PLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Optymalne modele PLS dla danych zawierających grupy uzyskane za pomocą algorytmu Kennarda i Stone'a oraz algorytmu Dupleks charakteryzowane były przez następujące wartości pierwiastka średniego błędu kwadratowego:
$RMSE_{(KS)} = 0{,}23$;
$RMSEP_{(KS)} = 0{,}37$
oraz
$RMSE_{(DU)} = 0{,}27$;
$RMSEP_{(DU)} = 0{,}35$.

Sieci neuronowe

Dane poddano kompresji, w wyniku której oryginalne zmienne zastąpiono czynnikami głównymi oraz wybranymi zmiennymi istotnymi. Czynniki główne obliczono stosując metodę PCA, a zmienne istotne pochodzą z modelu CART. Zmienne poddano skalowaniu do przedziału <-1, 1>.

Sieć neuronowa zawierała we wszystkich węzłach warstwy ukrytej funkcję typu tangens hiperboliczny, natomiast w węźle warstwy wyjściowej funkcję liniową. Jako pierwszy modelowany zestaw danych użyto pięciu czynników głównych (PCs) opisujących 99,02% wariancji danych. Próbki podzielono na zbiory za pomocą

algorytmu Kennarda i Stone'a. Optymalna sieć zawierała pięć węzłów wejściowych, cztery węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej. Sieć ta pozwoliła na przewidzenie zawartości wilgoci w próbkach z następującymi błędami:
$RMSE_{(KS/5PCs)} = 0{,}79$;
$RMSEP_{(KS/5PCs)} = 0{,}60$.

Kolejny zestaw danych zawierał zmienne istotne (ZM: 1, 36, 44, 76, 81, 96, 99, 161, 169, 207) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Optymalny model to sieć zawierająca dziesięć węzłów wejściowych oraz po jednym w warstwie ukrytej oraz wyjściowej. Optymalny model pozwolił na przewidzenie modelowanej własności z następującymi błędami:
$RMSE_{(KS/10ZM)} = 1{,}03$;
$RMSEP_{(KS/10ZM)} = 0{,}40$.

Następny modelowany zestaw danych to pięć czynników głównych (PCs) opisujących 99,02% wariancji danych, dla których obiekty podzielono na zbiory za pomocą algorytmu Duplex. Optymalna sieć zawierała pięć węzłów wejściowych, trzy węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej i pozwoliła na przewidzenie zawartości wilgoci w próbkach pszenicy:
$RMSE_{(DU/5PCs)} = 0{,}78$;
$RMSEP_{(DU/5PCs)} = 0{,}49$.

Ostatni zestaw danych zawierał zmienne istotne (ZM: 29, 37, 62, 64, 162, 167, 184, 224, 239) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Optymalny model to sieć zawierająca dziewięć węzłów wejściowych, trzy węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej. Pozwoliła ona na przewidzenie modelowanej własności z następującymi błędami:
$RMSE_{(DU/9ZM)} = 0{,}82$;
$RMSEP_{(DU/9ZM)} = 0{,}44$.

Neuronowe systemy rozmyte

Na koniec modelowano dane z zastosowaniem neuronowych systemów rozmytych. Skonstruowano modele NFS typu Sugeno pierwszego rzędu do modelowania danych zawierających zbiory uzyskane algorytmem Kennarda i Stone'a (KS) oraz algorytmem Duplex (DU).

Jako pierwszy modelowany zestaw danych użyto pięciu czynników głównych (PCs) opisujących 99,02% wariancji danych, dla których obiekty podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Iteracyjne uczenie modelu odbywało się w oparciu o metodę hybrydową. Najlepszy model wykorzystywał kratkowy schemat podziału przestrzeni danych. W ramach tego modelu skonstruowano 243 reguły logiczne poprzez przypisanie trzech funkcji przynależności na każdą zmienną. Jednakże z uwagi na fakt, iż konstruowany model NFS dotyczył zbioru zawierającego sto obiektów został on uznany za zawodny. Jako optymalny dla tych danych postanowiono wybrać inny model NFS, w ramach którego zastosowano metodę

grupowania różnicowego (o promieniu 0,8) do podziału przestrzeni danych. Skonstruowano
trzy reguły logiczne, a model uczono w oparciu o metodę hybrydową. Skonstruowany model pozwolił na przewidzenie modelowanej własności z następującymi błędami:
$RMSE_{(KS/5PCs)} = 0,67$;
$RMSEP_{(KS/5PCs)} = 0,74$.

Drugi zestaw danych zawierał zmienne istotne (ZM: 1, 36, 44, 76, 81, 96, 99, 161, 169, 207) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Skonstruowany model obarczony był błędami:
$RMSE_{(KS/10ZM)} = 0,47$;
$RMSEP_{(KS/10ZM)} = 0,62$.

Następny modelowany zestaw danych to pięć czynników głównych (PCs) opisujących 99,02% wariancji danych. Próbki podzielono na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystywał metodę grupowania różnicowego (o promieniu 0,8) do podziału przestrzeni danych. W ramach tego modelu skonstruowano pięć reguł logicznych. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Skonstruowany model pozwolił na przewidzenie liczby oktanowej z następującymi błędami:
$RMSE_{(DU/5PCs)} = 0,67$;
$RMSEP_{(DU/5PCs)} = 0,65$.

Czwarty zestaw danych zawierał zmienne istotne (ZM: 29, 37, 62, 64, 162, 167, 184, 224, 239) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Iteracyjne uczenie modelu odbywało się w oparciu o wsteczną propagację błędu. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano sześć reguł logicznych. Model NFS obarczony był błędami:
$RMSE_{(DU/9ZM)} = 0,65$;
$RMSEP_{(DU/9ZM)} = 0,75$.

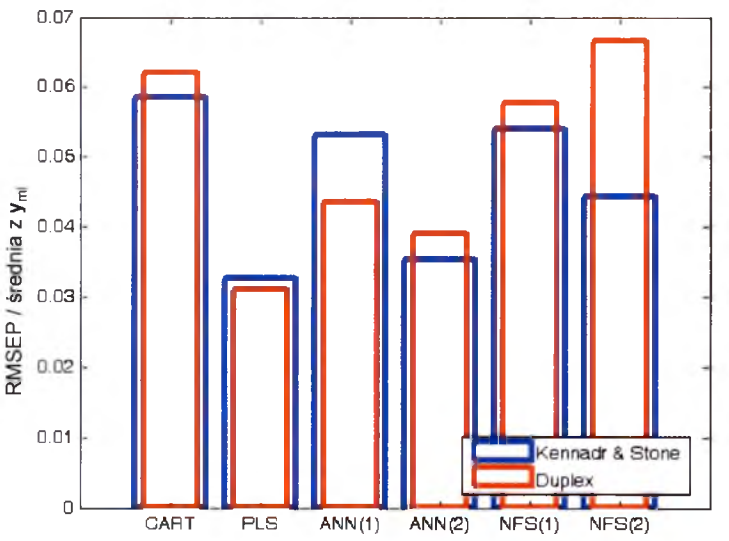
Podsumowanie

W poniższej tabeli zestawiono wyniki modelowania zawartości wilgoci w pszenicy (Tabela 4). Modele CART i PLS skonstruowano w oparciu o oryginalne zmienne. Natomiast do konstrukcji modeli ANN i NFS wykorzystano czynniki główne (PCs) i wybrane zmienne istotne (ZM). Wartości błędów RMSE oraz RMSEP obrazują odpowiednio odpasowanie modelu do danych i moc predykcyjną skonstruowanych modeli.

Tabela 4 Zestawienie wyników przeprowadzonych analiz dla modelowani zawartości wilgoci w pszenicy (Dane 3), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	RMSE	RMSEP	opis modelu
CART	KS	oryginalne	0,64	0,66	4 węzły terminalne
	DU	oryginalne	0,65	0,70	3 węzły terminalne
PLS	KS	oryginalne	0,23	0,37	8 czynników ukrytych
	DU	oryginalne	0,27	0,35	7 czynników ukrytych
ANN	KS	5 PCs	0,79	0,60	4 węzły w warstwie ukrytej
		10 ZM	1,03	0,40	1 węzeł w warstwie ukrytej
	DU	5 PCs	0,78	0,49	3 węzły w warstwie ukrytej
		9 ZM	0,82	0,44	3 węzły w warstwie ukrytej
NFS	KS	5 PCs	0,67	0,74	3 reguły logiczne
		10 ZM	0,47	0,62	2 reguły logiczne
	DU	5 PCs	0,67	0,65	5 reguł logicznych
		9 ZM	0,65	0,75	6 reguł logicznych

Rys. 55 przedstawia porównanie otrzymanych wyników modelowania za pomocą zastosowanych metod (Tabela 4, piąta kolumna). Wartości błędu zostały podzielone przez wartość średnią zmiennej zależnej ze zbioru modelowego celem porównania wyników dla różnych zestawów danych.

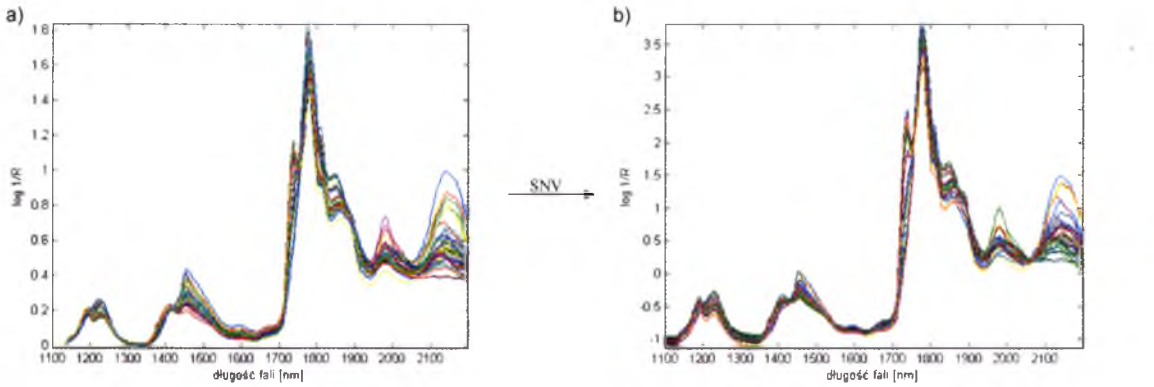


Rys. 55 Wykres wartości pierwiastka średniego błędu kwadratowego charakteryzujący konstruowane modele celem modelowania zawartości wilgoci w próbkach pszenicy, gdzie indeksy oznaczają modele konstruowane w oparciu odpowiednio o dane zawierające (1) czynniki główne oraz (2) zmienne istotne

W przypadku modeli konstruowanych dla danych zawierających zbiory z algorytmu Kennarda i Stone'a otrzymane wyniki dla modelu NFS (konstruowanego w oparciu o czynniki główne) były porównywalne z wynikami dla modelu CART oraz modelu ANN (konstruowanego w oparciu o czynniki główne). Mniejsze wartości błędu otrzymano dla modelu NFS (konstruowanego w oparciu o wybrane zmienne), które są porównywalne z wynikami dla modelu CART oraz modelu ANN (konstruowanego w oparciu o wybrane zmienne). Taka zależność nie występuje w przypadku modeli konstruowanych w oparciu o dane zawierające zbiory utworzone algorytmem Duplex.

9.4 Dane 4: Modelowanie liczby grup -OH w cząsteczkach polioli

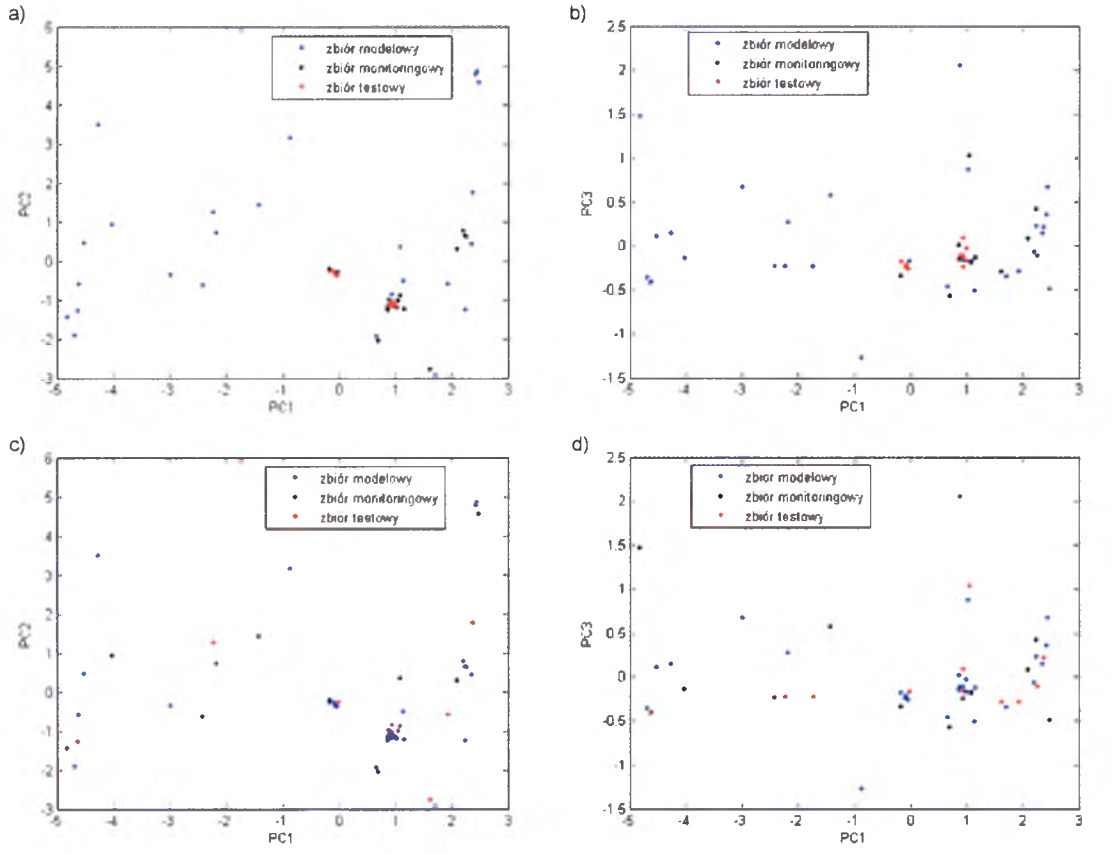
Dla 52 próbek słodzików będących alkoholami polihydroksylowymi zarejestrowano widma w bliskiej podczerwieni [107]. Widma rejestrowano w wariancie odbiciowym spektroskopii NIR (log 1/R) w zakresie od 1100 nm do 2200 nm. W postaci wektora zmiennej zależnej zapisano liczbę grup hydroksylowych dla każdej próbki. Dane o wymiarowości 52 x 512 poddano wstępnej obróbce stosując transformację SNV (Rys. 56).



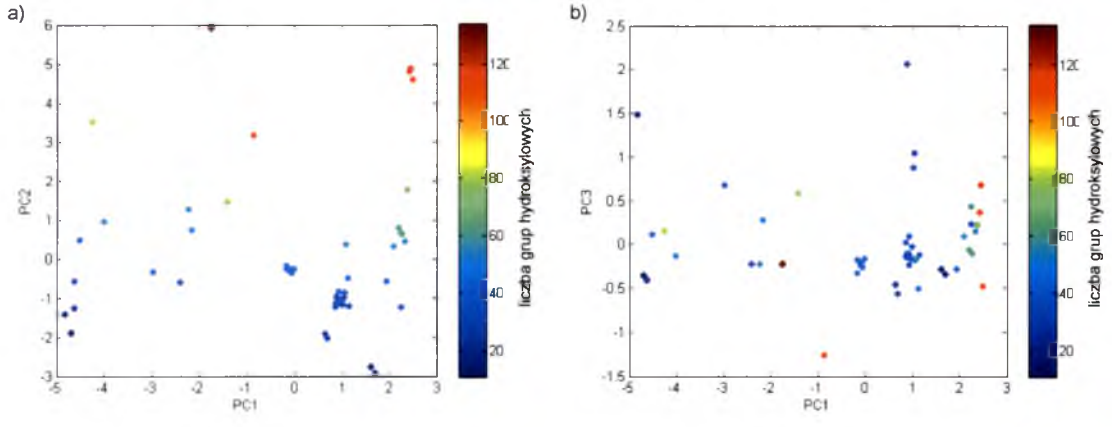
Rys. 56 Widma NIR 52 próbek słodzików a) przed i b) po transformacji SNV

Eksploracja i przygotowanie danych

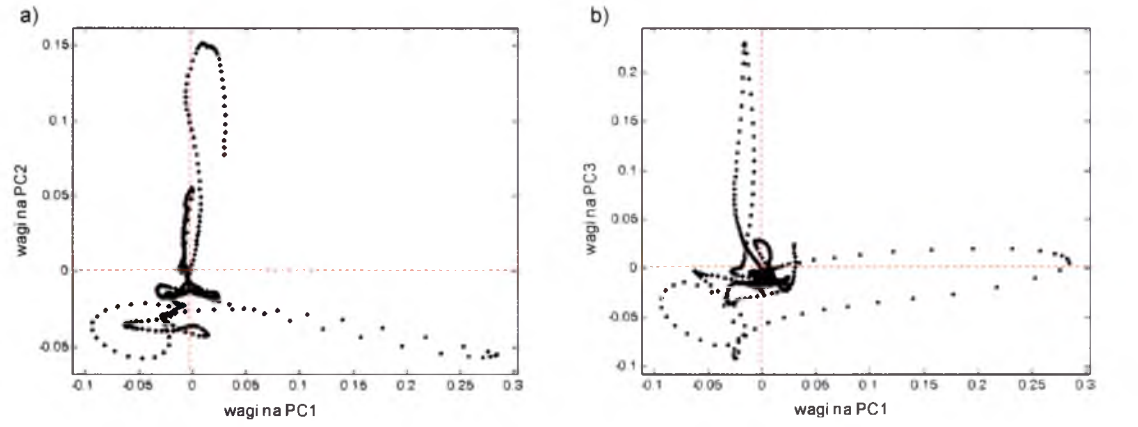
Analiza czynników głównych pozwoliła na wizualizację i eksplorację danych (Rys. 57-60). W toku analizy nie wykryto obiektów odległych. Ponadto na rysunku 58a widoczne jest, że liczba grup hydroksylowych w cząsteczkach alkoholi polihydroksylowych jest dodatnio skorelowana z drugim czynnikiem głównym (PC2).



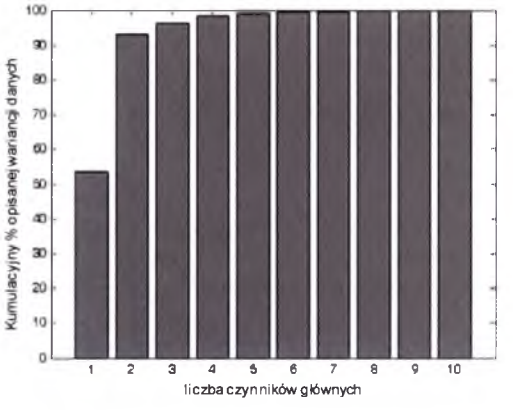
Rys. 57 Projekcja 52 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy oraz drugi czynnik główny oraz b, d) przez pierwszy oraz trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 58 Projekcja obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono liczbę grup hydroksylowych w cząsteczkach polioli



Rys. 59 Projekcja parametrów na płaszczyznę wag zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny



Rys. 60 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

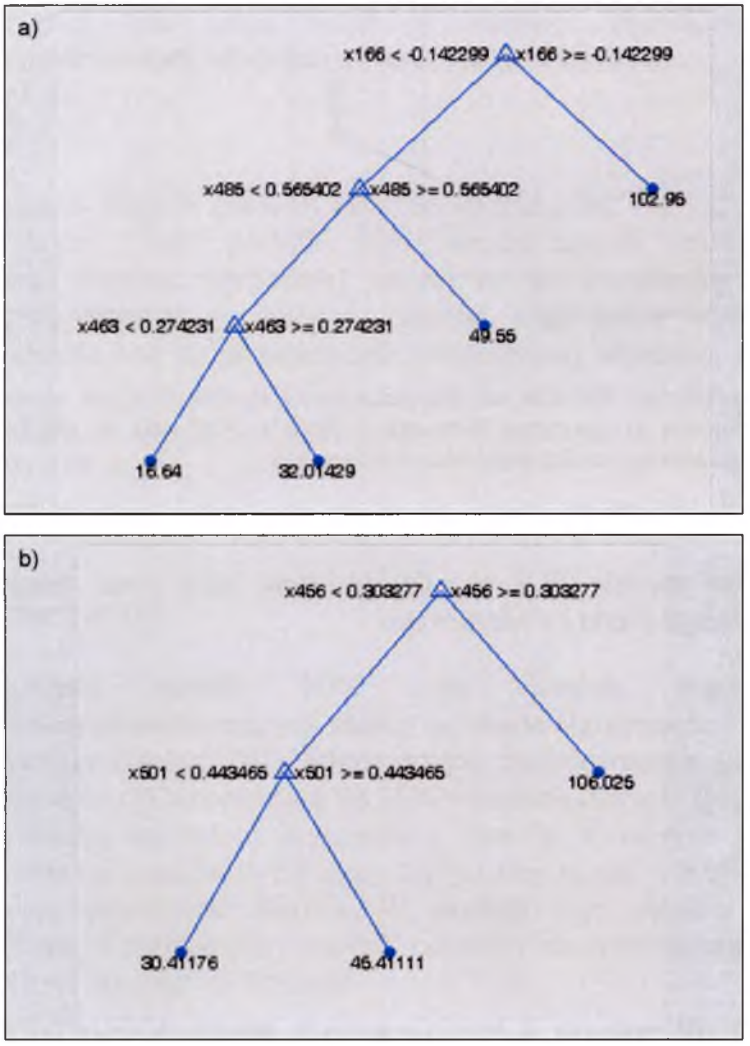
W dalszej kolejności 52 próbki słodzików podzielono na trzy zbiory przypisując 30 obiektów do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$) oraz po 11 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Zmienna zależna dla wszystkich zbiorów ($\mathbf{y}_{ml}$, $\mathbf{y}_{mr}$, $\mathbf{y}_{tt}$) została poddana centrowaniu. Tak utworzone zbiory zostały poddane modelowaniu metodą CART oraz PLS.

Drzewa klasyfikacji i regresji

Jako pierwszy skonstruowano model drzew klasyfikacji i regresji (CART). Optymalne binarne drzewo decyzyjne skonstruowane w oparciu o zbiory tworzone za pomocą algorytmu Kennarda i Stone'a miało cztery węzły terminalne (Rys. 61a). Zmienne wskazane w modelu jako decyzyjne to zmienna 166, 463 i 485 oraz zmienna

310 wskazana przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły odpowiednio:
$RMSE_{(KS)} = 11{,}50$;
$RMSEP_{(KS)} = 6{,}01$.

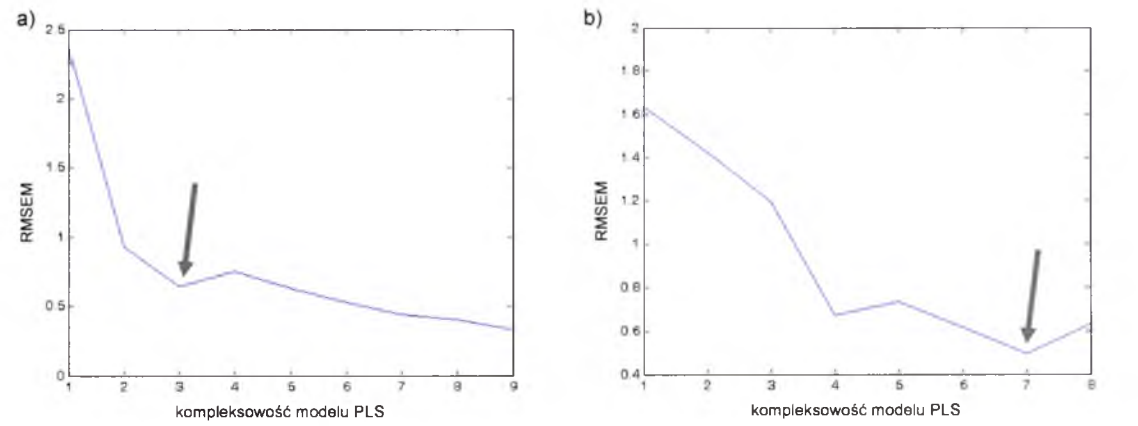
Natomiast optymalny model skonstruowany w oparciu o dane zawierające zbiory utworzone za pomocą algorytmu Duplex miał trzy węzły terminalne (Rys. 61b). Zmienne wskazane w modelu jako decyzyjne to zmienna 456 i 501 oraz zmienne 18, 185 i 445 wskazane przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły:
$RMSE_{(DU)} = 8{,}68$;
$RMSEP_{(DU)} = 14{,}88$.



Rys. 61 Optymalne drzewo CART skonstruowane celem modelowania liczby grup hydroksylowych w cząsteczkach alkoholi polihydroksylowych dla zbiorów utworzonych za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU)

Metoda częściowych najmniejszych kwadratów

Drugą wykorzystaną techniką modelowania danych była metoda częściowych najmniejszych kwadratów. W oparciu o zbiór monitoringowy wyznaczono kompleksowość modelu PLS. Wybrano trzy czynniki ukryte do konstrukcji optymalnego modelu konstruowanego w oparciu o dane zawierające zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 62a) oraz siedem czynników ukrytych dla modelowania danych zawierających zbiory otrzymane za pomocą algorytmu Duplex (DU, Rys. 62b).



Rys. 62 Wykres zależności RMSEM od kompleksowości modelu PLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Optymalne modele PLS charakteryzowane były przez następujące wartości pierwiastka średniego błędu kwadratowego:
$RMSE_{(KS)} = 2{,}56$;
$RMSEP_{(KS)} = 2{,}01$
oraz
$RMSE_{(DU)} = 1{,}46$;
$RMSEP_{(DU)} = 2{,}15$.

Sieci neuronowe

Przed modelowaniem z wykorzystaniem metod ANN i NFS dane poddano redukcji i skalowaniu do przedziału <-1, 1>. Konstruowana sieć ANN zawierała w węzłach warstwy ukrytej funkcje typu tangens hiperboliczny, a w węźle warstwy wyjściowej funkcję liniową. Jako pierwszy modelowany zestaw danych użyto cztery czynniki główne (PCs) opisujące 98,35 % wariancji danych, dla których obiekty podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalna sieć zawierała cztery węzły wejściowe, cztery węzły w warstwie ukrytej i ukrytej oraz jeden węzeł w warstwie wyjściowej. Sieć ta pozwoliła na przewidzenie liczby grup hydroksylowych z następującymi błędami:

$RMSE_{(KS/4PCs)} = 11,46$;
$RMSEP_{(KS/4PCs)} = 1,28$.

Kolejny zestaw danych zawierał zmienne istotne (ZM: 106, 310, 463, 485) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Optymalny model to sieć zawierająca cztery węzły wejściowe, trzy w warstwie ukrytej oraz jeden w warstwie wyjściowej. Pozwoliła ona na przewidzenie modelowanej własności z następującymi błędami:
$RMSE_{(KS/4ZM)} = 2,00$;
$RMSEP_{(KS/4ZM)} = 0,92$.

Następny modelowany zestaw danych to cztery czynniki główne (PCs) opisujące 98,35% wariancji danych, dla których próbki podzielono na zbiory za pomocą algorytmu Duplex. Optymalna sieć zawierała cztery węzły wejściowe, trzy węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej i pozwoliła na przewidzenie modelowanej własności z następującymi błędami:
$RMSE_{(DU/4PCs)} = 1,92$;
$RMSEP_{(DU/4PCs)} = 1,95$.

Ostatni zestaw danych zawierał zmienne istotne (ZM: 18, 185, 445, 456, 501) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Optymalny model to sieć zawierająca pięć węzłów wejściowych, cztery węzły w warstwie ukrytej oraz jeden węzeł w warstwie wyjściowej. Pozwoliła ona na przewidzenie modelowanej własności z następującymi błędami:
$RMSE_{(DU/5ZM)} = 1,54$;
$RMSEP_{(DU/5ZM)} = 1,34$.

Neuronowe systemy rozmyte

Skonstruowano modele NFS typu Sugeno pierwszego rzędu do modelowania danych zawierających zbiory uzyskane algorytmem Kennarda i Stone (KS) oraz algorytmem Duplex (DU). Jako pierwszy modelowany zestaw danych użyto cztery czynniki główne (PCs) opisujące 98,35% wariancji danych. Obiekty podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Iteracyjne uczenie modelu odbywało się w oparciu o model hybrydowy. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano 23 reguły logiczne. Konstruowany model pozwolił na przewidzenie liczby grup hydroksylowych z następującymi błędami:
$RMSE_{(KS/4PCs)} = 0,00$;
$RMSEP_{(KS/4PCs)} = 2,38$.

Drugi zestaw danych zawierał zmienne istotne (ZM: 166, 310, 463, 485) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model wykorzystuje metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano 19 reguł logicznych. Konstruowany model obarczony był błędami:

$RMSE_{(KS/4ZM)} = 0,00;$
$RMSEP_{(KS/4ZM)} = 1,32.$

Następny modelowany zestaw danych to cztery czynniki główne (PCs) opisujące 98,35% wariancji danych, dla których obiekty podzielono na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Konstruowany model pozwolił na przewidzenie liczby grup hydroksylowych z następującymi błędami:
$RMSE_{(DU/4PCs)} = 1,54;$
$RMSEP_{(DU/4PCs)} = 2,44.$

Czwarty zestaw danych zawierał zmienne istotne (ZM: 18, 185, 445, 456, 501) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Iteracyjne uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Skonstruowany model obarczony był błędami:
$RMSE_{(DU/5ZM)} = 0,85;$
$RMSEP_{(DU/5ZM)} = 2,15.$

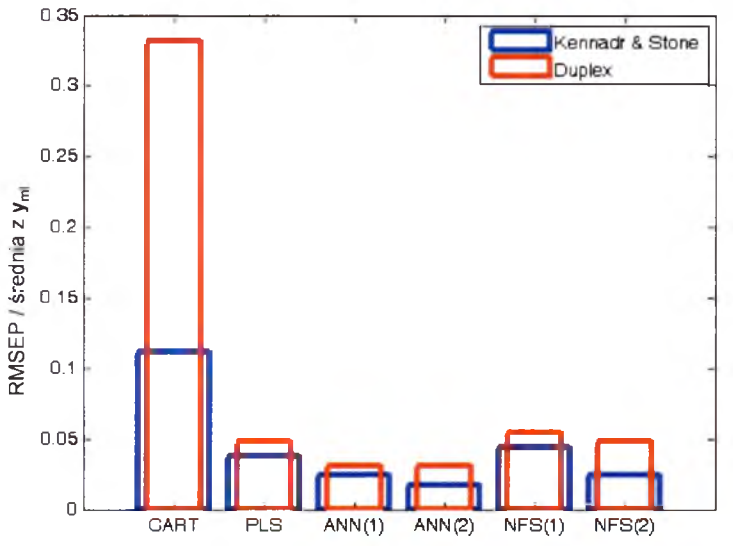
Podsumowanie

Tabela 5 zawiera wyniki modelowania liczby grup hydroksylowych w cząsteczkach alkoholi polihydroksylowych w postaci błędów RMSE oraz RMSEP. Modele skonstruowano w oparciu o oryginalne zmienne (CART i PLS), lub czynniki główne i wybrane zmienne istotne (ANN, NFS).

Tabela 5 Zestawienie wyników przeprowadzonych analiz dla modelowania liczby grup hydroksylowych w cząsteczkach alkoholi polihydroksylowych (Dane 4)

model	algorytm tworzenia zbiorów	modelowane zmienne	RMSE	RMSEP	opis modelu
CART	KS	oryginalne	11,50	6,01	4 węzły terminalne
	DU	oryginalne	8,68	14,88	3 węzły terminalne
PLS	KS	oryginalne	2,56	2,01	3 czynniki ukryte
	DU	oryginalne	1,46	2,15	7 czynników ukrytych
ANN	KS	4 PCs	11,46	1,28	4 węzły w warstwie ukrytej
		4 ZM	2,00	0,92	3 węzły w warstwie ukrytej
	DU	4 PCs	1,92	1,95	3 węzły w warstwie ukrytej
		5 ZM	1,54	1,34	4 węzły w warstwie ukrytej
NFS	KS	4 PCs	0,00	2,38	23 reguły logiczne
		4 ZM	0,00	1,32	19 reguł logicznych
	DU	4 PCs	1,54	2,44	2 reguły logiczne
		5 ZM	0,85	2,15	2 reguły logiczne

Na rysunku 63 przedstawiono wykres porównujący otrzymane wyniki modelowania za pomocą zastosowanych metod. Wartości błędu zostały podzielone przez wartość średnią zmiennej zależnej ze zbioru modelowego.

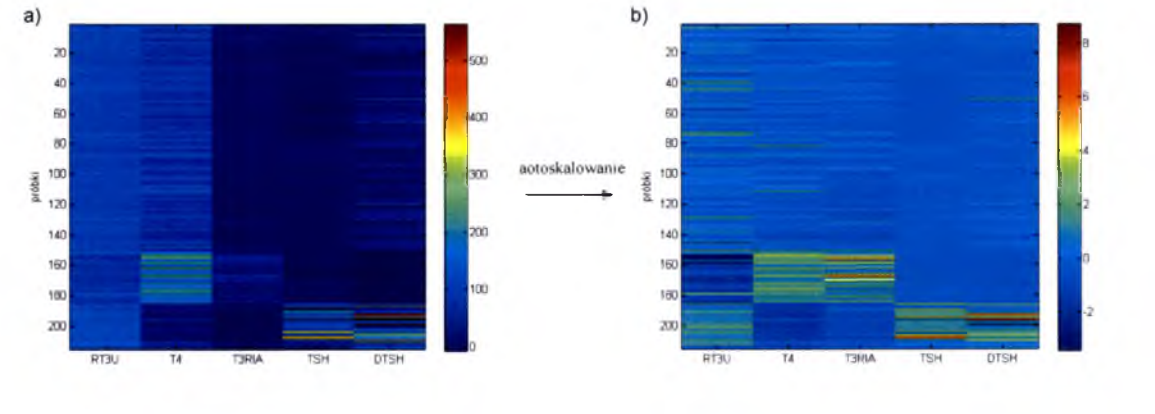


Rys. 63 Wykres wartości pierwiastka średniego błędu kwadratowego charakteryzujący konstruowane modele celem modelowania liczby grup hydroksylowych w cząsteczkach słodzików, gdzie indeksy oznaczają modele konstruowane w oparciu odpowiednio o dane zawierające (1) czynniki główne oraz (2) zmienne istotne

Metoda NFS pozwoliła na konstrukcję modeli obarczonych mniejszym błędem niż metoda CART, jednocześnie dostarczając reguł logicznych. Ponadto metoda NFS dostarczyła modeli o porównywalnej mocy predykcyjnej do powszechnie stosowanych metod PLS oraz ANN.

9.5 Dane 5: Modelowanie prawidłowego funkcjonowania tarczycy

Przebadano 215 osób pod kątem anomalii w funkcjonowaniu tarczycy [108]. Wśród badanych próbek znalazło się 65 próbek tkanki należących do osób, których tarczyca nie funkcjonowała prawidłowo (35 osób z nadczynnością gruczołu oraz 30 z jego niedoczynnością). Grupa kontrolna liczyła 150 osób. W ramach badań próbki tkanki tarczycy poddano analizie za pomocą pięciu testów przeprowadzonych *in vitro*: RT3U, T4, T3RIA, TSH, DTSH (Rys. 64a).

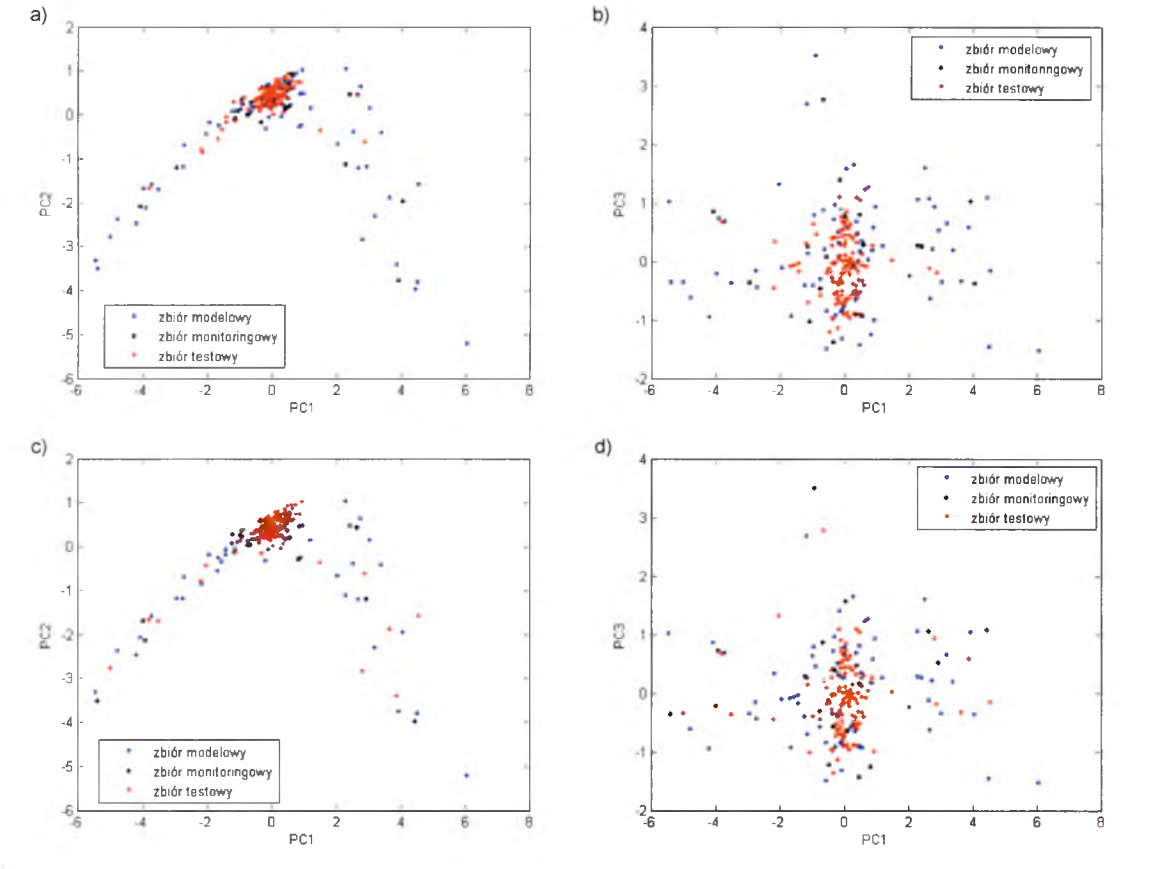


Rys. 64 Wyniki pięciu testów prawidłowego funkcjonowania tarczycy przeprowadzone na 215 pacjentach a) przed i b) po autoskalowaniu

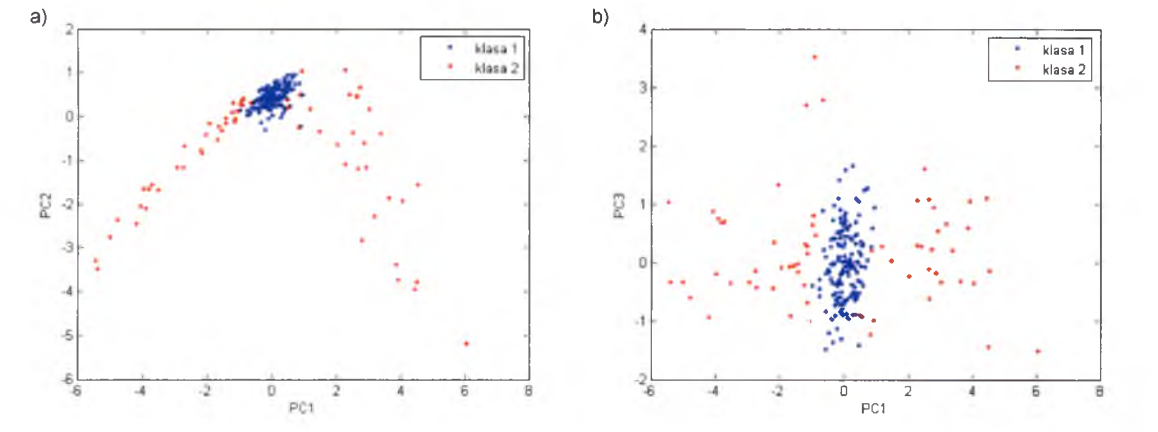
Przynależność obiektów do jednej z dwóch grup została zakodowana w postaci binarnej zmiennej zależnej: klasa 1: tkanki zdrowych pacjentów (150 obiektów), klasa 2 próbki od chorych osób (65 obiektów). Analizowane dane miały wymiarowość 215 x 5, zostały poddane autoskalowaniu celem zrównoważenia istotności poszczególnych parametrów (Rys. 64b).

Eksploracja i przygotowanie danych

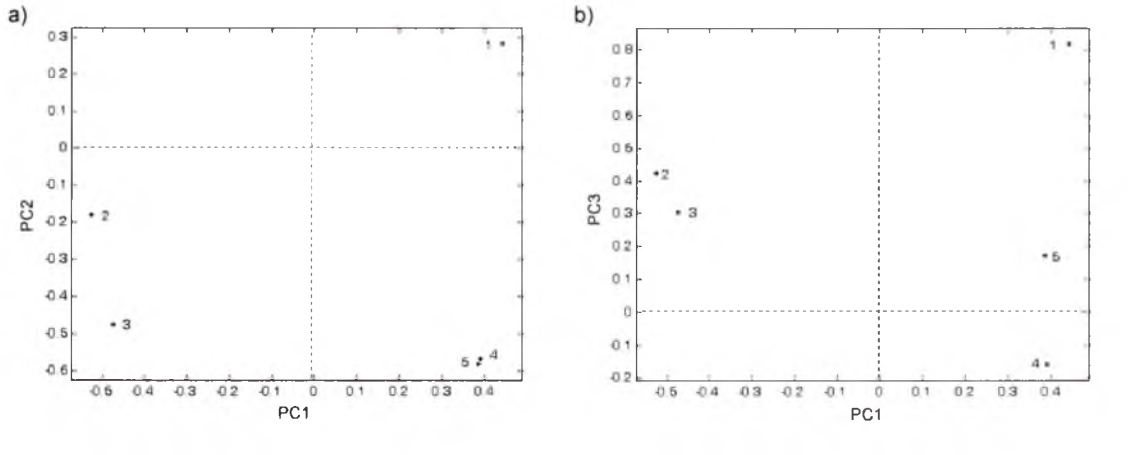
Zastosowano analizę czynników głównych celem wizualizacji i eksploracji danych (Rys. 65-68). W toku analizy nie wykryto obiektów odległych. Zaznaczenie stanu zdrowia pacjenta na projekcji obiektów na płaszczyznę zdefiniowana przez czynniki główne pozwoliło odnotować fakt, iż klasa 2 wykazuje się dużo większą wariancją wewnątrzgrupową niż klasa 1 (Rys. 66). Ponadto projekcja parametrów na płaszczyznę czynników głównych pokazała, że istnieje korelacja pomiędzy parametrami 2 i 3 oraz 4 i 5 (Rys. 67).



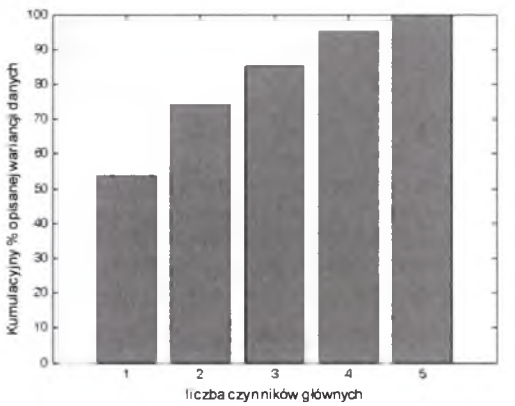
Rys. 65 Projekcja 215 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 66 Projekcja 215 obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono stan zdrowia pacjenta (klasa 1 – osoby zdrowe, klasa 2 – osoby chore)



Rys. 67 Projekcja parametrów na płaszczyznę zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny, gdzie: 1 – RT3U, 2 – T4, 3 – T3RIA, 4 – TSH, 5 – DTSH



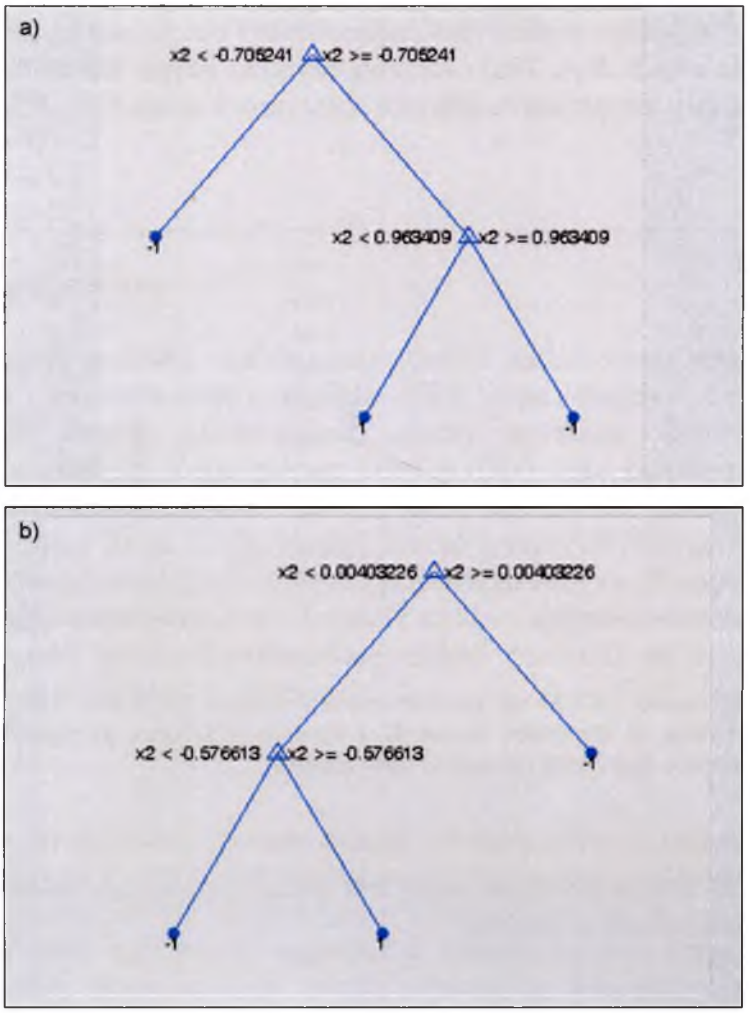
Rys. 68 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

Dane podzielono na trzy zbiory przypisując po 40 obiektów z każdej klasy do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$) oraz po 10 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz resztę (100 obiektów z klasy 1/zdrowych oraz 15 z klasy 2/chorych) do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Zmienna zależna dla wszystkich zbiorów ($\mathbf{y}_{ml}$, $\mathbf{y}_{mr}$, $\mathbf{y}_{tt}$) została zakodowana binarnie. Tak utworzone zbiory zostały poddane analizie metodą CART oraz PLS.

Drzewa klasyfikacji i regresji

Pierwszą wykorzystaną metodą była metoda CART. Optymalne binarne drzewo decyzyjne skonstruowane w oparciu o zbiory tworzone za pomocą algorytmu Kennarda i Stone'a miało trzy węzły terminalne (Rys. 69a). Model dwukrotnie wskazał zmienną

numer dwa (T4) jako decyzyjną. Procentowe wartości poprawnie sklasyfikowanych próbek przez model wyniosły odpowiednio:
$CCR_{(KS)} = 99{,}00\%$;
$CCRT_{(KS)} = 92{,}17\%$.

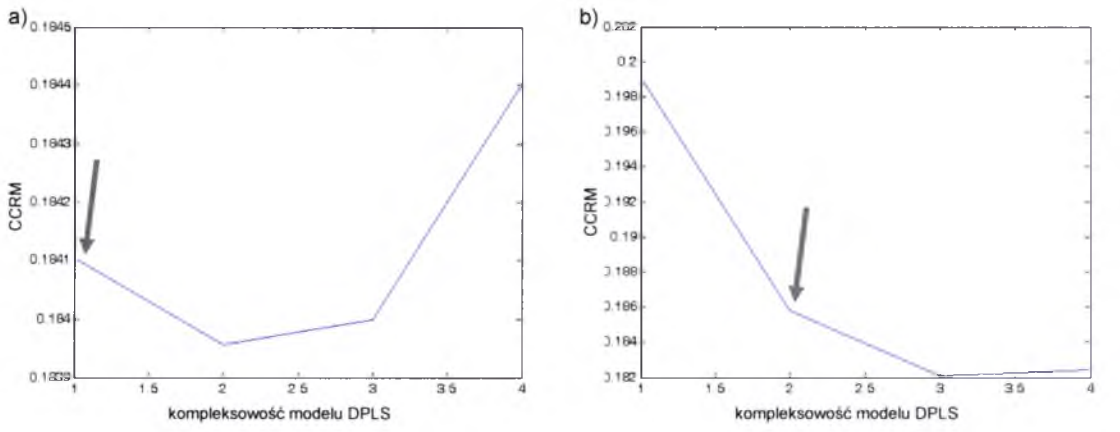


Rys. 69 Optymalne drzewo CART skonstruowane celem klasyfikacji pacjentów zdrowych oraz z dysfunkcją tarczycy w oparciu o zbiory utworzone za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie (1) klasa 1 i (-1) klasa 2

Optymalny model skonstruowany w oparciu o dane zawierające zbiory utworzone za pomocą algorytmu Duplex miał także trzy węzły terminalne (Rys. 69b) i dwukrotnie wskazał zmienną numer dwa (T4) jako decyzyjną. Procentowe wartości poprawnie sklasyfikowanych próbek przez model wyniosły:
$CCR_{(DU)} = 95{,}00\%$;
$CCRT_{(DU)} = 92{,}17\%$.

Metoda częściowych najmniejszych kwadratów w wariancie dyskryminacyjnym

Jako drugą zastosowano metodę częściowych najmniejszych kwadratów w wariancie dyskryminacyjnym (DPLS) [109]. W oparciu o zbiór monitoringowy wyznaczono kompleksowość modelu DPLS. Z uwagi na niewielkie różnice pomiędzy wartościami CCRM wybrano jeden czynnik ukryty do konstrukcji optymalnego modelu konstruowanego w oparciu o dane zawierające zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 70a) oraz dwa czynniki ukryte dla modelowania danych zawierających zbiory otrzymane za pomocą algorytmu Duplex (DU, Rys. 70b).



Rys. 70 Wykres zależności CCRM od kompleksowości modelu DPLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Finalny model charakteryzowany był przez następujące procentowe wartości poprawnie sklasyfikowanych próbek:
$CCR_{(KS)} = 81{,}25\%$;
$CCRT_{(KS)} = 91{,}30\%$
oraz
$CCR_{(DU)} = 77{,}30\%$;
$CCRT_{(DU)} = 96{,}52\%$.

Sieci neuronowe

Z uwagi na małą wymiarowość danych modele ANN oraz NFS zostały skonstruowane w oparciu o oryginalne zmienne. Zmienne poddano skalowaniu do przedziału od -1 do 1. Model ANN w swej sieci zawierał w węzłach obydwu warstw funkcję typu tangens hiperboliczny. Jako pierwszy modelowany zestaw danych użyto obiektów podzielonych na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalna sieć zawierała pięć węzłów wejściowych i po jednym węźle w warstwie ukrytej oraz wyjściowej. Sieć ta pozwoliła na przewidzenie stanu zdrowia pacjenta z następującym sukcesem:

$CCR_{(KS)} = 91,25\%$;
$CCRT_{(KS)} = 97,39\%$.

Kolejny zestaw danych zawierał obiekty przydzielone do zbiorów przez algorytm Duplex. Optymalny model skonstruowany dla tych danych także zawierał pięć węzłów wejściowych i po jednym węźle w warstwie ukrytej oraz wyjściowej. Sieć ta pozwoliła na przewidzenie modelowanej własności z następującym powodzeniem:
$CCR_{(DU)} = 87,50\%$;
$CCRT_{(DU)} = 98,26\%$.

Neuronowe systemy rozmyte

Jako ostatnią technikę modelowania danych zastosowano neuronowe systemy rozmyte (NFS). Skonstruowano modele NFS typu Sugeno pierwszego rzędu do modelowania danych zawierających zbiory uzyskane zarówno algorytmem Kennarda i Stone (KS) jak i algorytmem Duplex (DU). Jako pierwszy skonstruowano model dla danych zawierających obiekty, które podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalny model wykorzystywał metodę grupowania różnicowego (o promieniu 1,0) do podziału przestrzeni danych. W ramach tego modelu skonstruowano trzy reguły logiczne. Uczenie modelu odbywało się z zastosowaniem wstecznej propagacji błędu. Konstruowany model pozwolił na przewidzenie stanu zdrowia pacjenta z następującym sukcesem:
$CCR_{(KS)} = 97,50\%$;
$CCRT_{(KS)} = 99,13\%$.

Następny modelowany zestaw danych zawiera obiekty podzielone na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystywał metodę grupowania różnicowego (o promieniu 0,4) do podziału przestrzeni danych. W ramach tego modelu skonstruowane zostały trzy reguły logiczne, a uczenie modelu odbywało się według metody hybrydowej. Skonstruowany model pozwolił na przewidzenie stanu pacjenta z następującym sukcesem:
$CCR_{(DU)} = 97,50\%$;
$CCRT_{(DU)} = 99,12\%$.

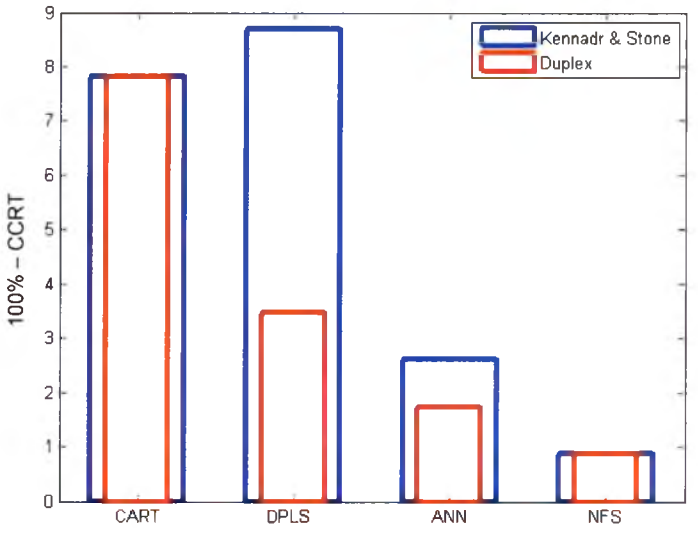
Podsumowanie

Poniższa tabela zawiera wyniki modelowania prawidłowego funkcjonowania tarczycy. Parametry charakteryzujące jakość skonstruowanych modeli to procent poprawnie sklasyfikowanych próbek ze zbioru modelowego (CCR) oraz z niezależnego zbioru testowego (CCRT). Błędy te charakteryzują odpowiednio odpasowanie modelu do danych i moc predykcyjną skonstruowanych modeli. Wszystkie modele zostały skonstruowane w oparciu o oryginalne zmienne.

Tabela 6 Zestawienie wyników przeprowadzonych analiz dla modelowania prawidłowego funkcjonowania tarczycy (Dane 5), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	CCR [%]	CCRT [%]	opis modelu
CART	KS	oryginalne	99,00	92,17	3 węzły terminalne
	DU	oryginalne	95,00	92,17	3 węzły terminalne
DPLS	KS	oryginalne	81,25	91,30	1 czynnik ukryty
	DU	oryginalne	77,30	96,52	2 czynniki ukryte
ANN	KS	oryginalne	91,25	97,39	1 węzeł w warstwie ukrytej
	DU	oryginalne	87,50	98,26	1 węzeł w warstwie ukrytej
NFS	KS	oryginalne	97,50	99,13	3 reguły logiczne
	DU	oryginalne	97,50	99,12	3 reguły logiczne

Podsumowanie wyników w postaci wykresu procentu błędnie sklasyfikowanych próbek za pomocą zastosowanych metod przedstawiono na rysunku 71. Wyniki dla modelu CART były gorsze od wyników dla pozostałych metod modelowania danych. Technika NFS pozwoliła na konstrukcję modeli o największej mocy predykcyjnej w porównaniu zarówno do metod DPLS i ANN.



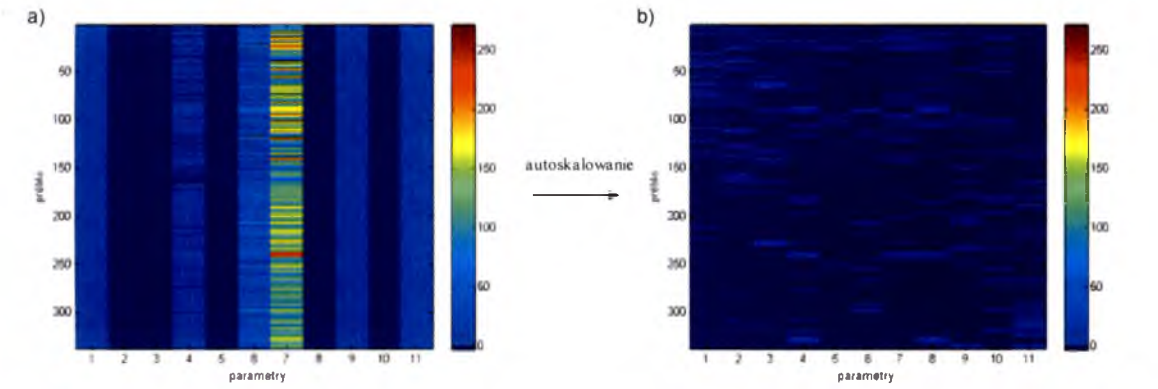
Rys. 71 Wykres procentu błędnie sklasyfikowanych próbek (100% – CCR) charakteryzujący konstruowane modele

9.6 Dane 6: Modelowanie jakości win białych

Analizowano jakość portugalskiego wina *vinho verde* [110], dla którego zmierzono następujące parametry: kwasowość trwała, kwasowość przemijająca oraz zawartość kwasu cytrynowego, pozostałości cukru, chlorków, wolnego dwutlenku siarki, całkowita zawartość tlenku siarki, gęstość, pH, zawartość siarczanów i alkoholu. Do modelowania wybrano dwa zestawy próbek. Pierwszy zestaw A zawierał dwie grupy próbek, jedną o niskiej jakości (163 próbki) oraz drugą o wysokiej jakości (175 próbek). Parametry jakości dla obydwu grup określono w dziesięciopunktowej skali i wynosiły one odpowiednio dla grupy pierwszej oraz drugiej: 4 i 8. Drugi zestaw danych B zawierał także dwa zbiory obiektów (grupa pierwsza 1457 próbek, grupa druga 2198 próbek), jednakże w tym przypadku różnica w jakości wina była nieznaczna która wynosiła 5 dla grupy pierwszej i 6 dla drugiej. Wyniki analizy zestawu A oraz B przedstawione są kolejno poniżej.

Zestaw A

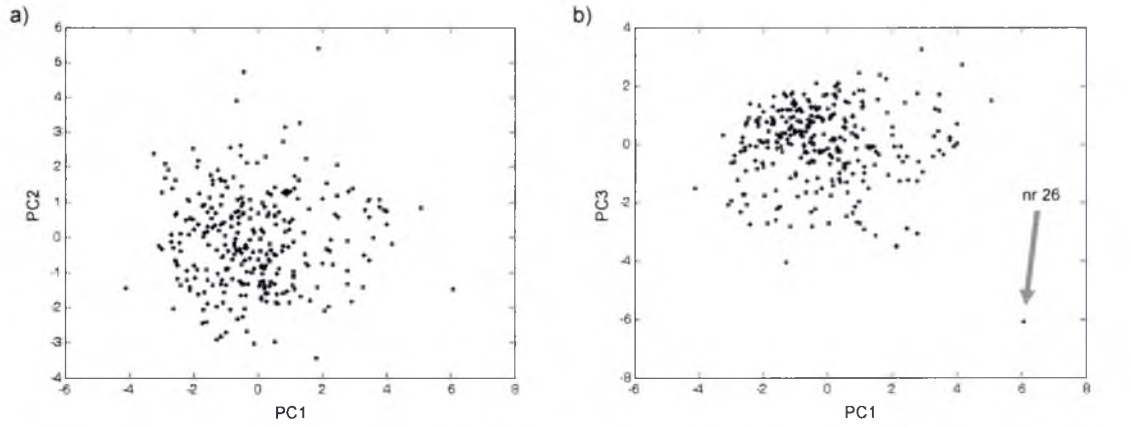
Dane z zestawu A miały wymiarowość 338 **x** 11. Przynależność do klas zakodowano w postaci binarnej zmiennej zależnej **y**. Dane poddano autoskalowaniu celem zrównoważenia wkładu poszczególnych parametrów (Rys. 72).



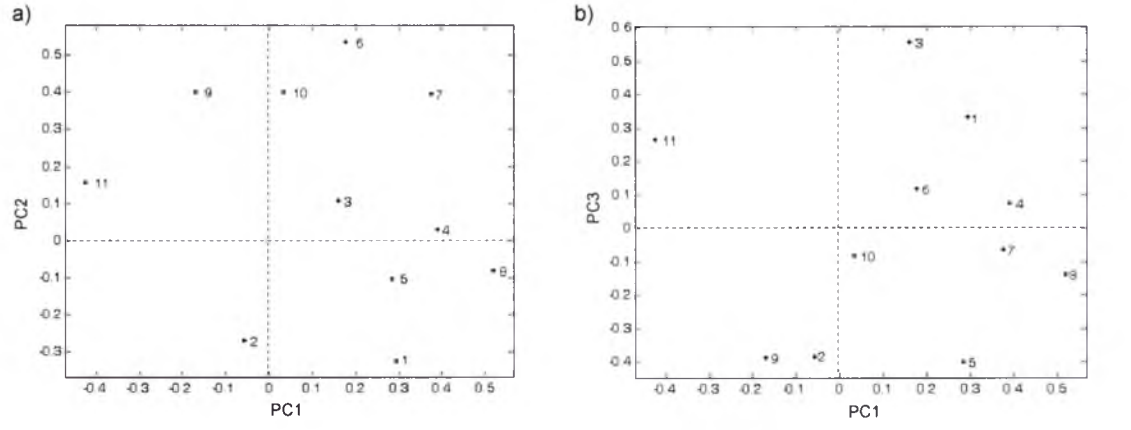
Rys. 72 Wartości jedenastu parametrów dla 338 próbek wina a) przed i b) po autoskalowaniu, gdzie: 1 – kwasowość trwała, 2 – kwasowość przemijająca, 3 – zawartość kwasu cytrynowego, 4 – pozostałości cukru, 5 – chlorki, 6 – wolny dwutlenek siarki, 7 – całkowita zawartość tlenku siarki, 8 – gęstość, 9 – pH, 10 – zawartość siarczanów i 11 – zawartość alkoholu

Eksploracja i przygotowanie danych (A)

Eksploracja danych z wykorzystaniem metody PCA ujawniła występowanie jednego obiektu odległego w analizowanych danych (Rys. 73). Projekcja parametrów na płaszczyznę zdefiniowaną przez czynniki główne nie pozwoliła stwierdzić faktu występowania silnie skorelowanych parametrów.

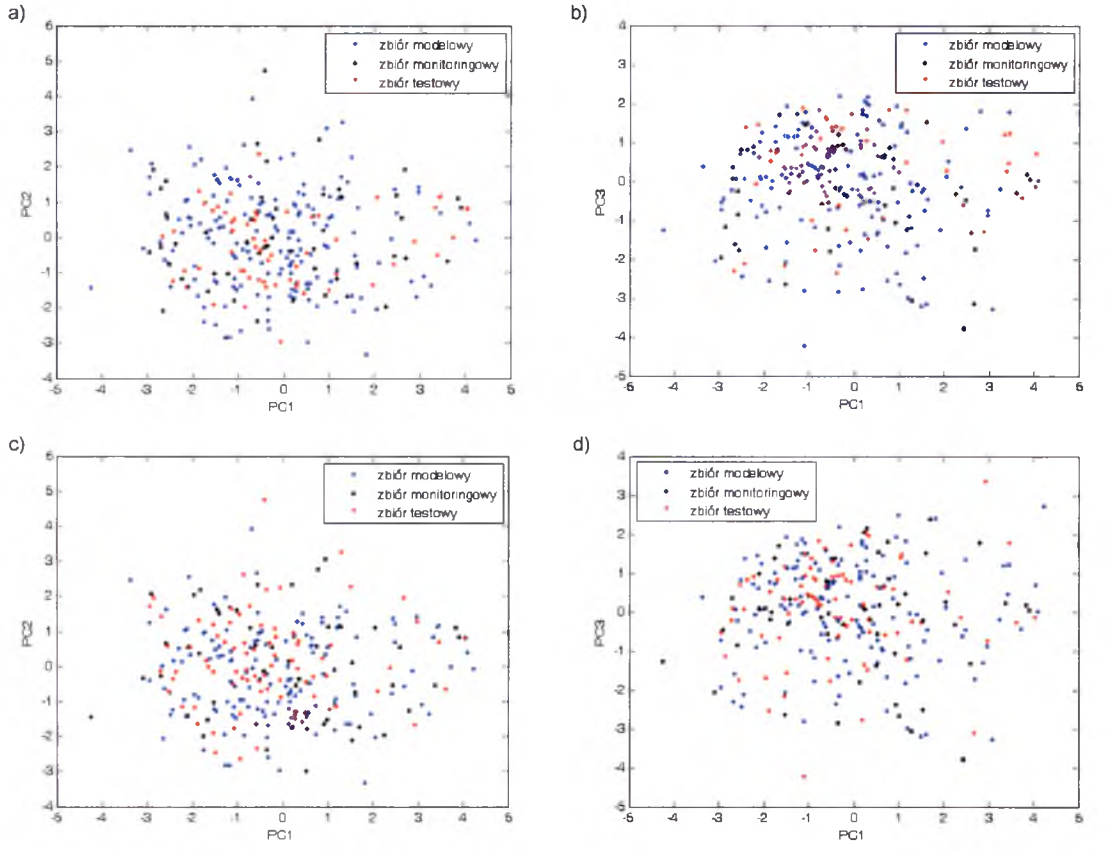


Rys. 73 Projekcja 338 obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono obiekt odległy

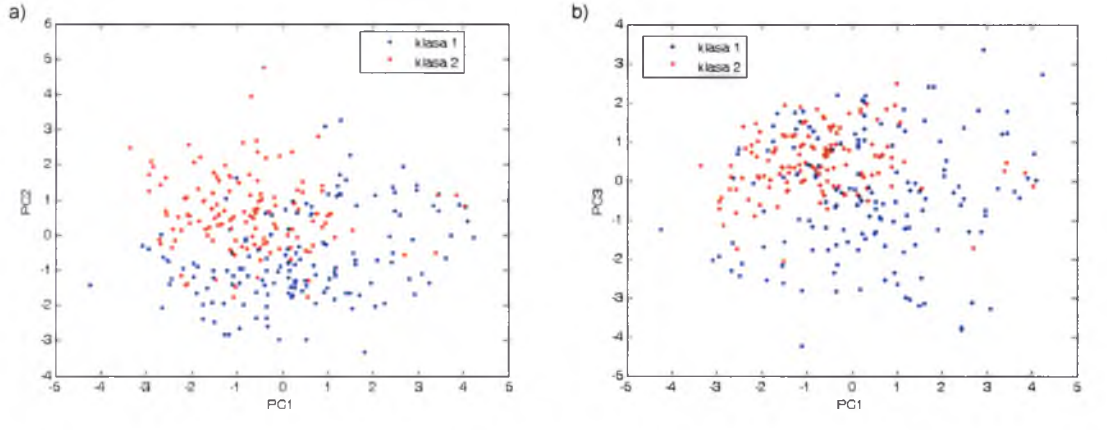


Rys. 74 Projekcja parametrów na płaszczyznę wag zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny, gdzie: 1 – kwasowość trwała, 2 – kwasowość przemijająca, 3 – zawartość kwasu cytrynowego, 4 – pozostałości cukru, 5 – chlorki, 6 – wolny dwutlenek siarki, 7 – całkowita zawartość tlenku siarki, 8 – gęstość, 9 – pH, 10 – zawartość siarczanów i 11 – zawartość alkoholu

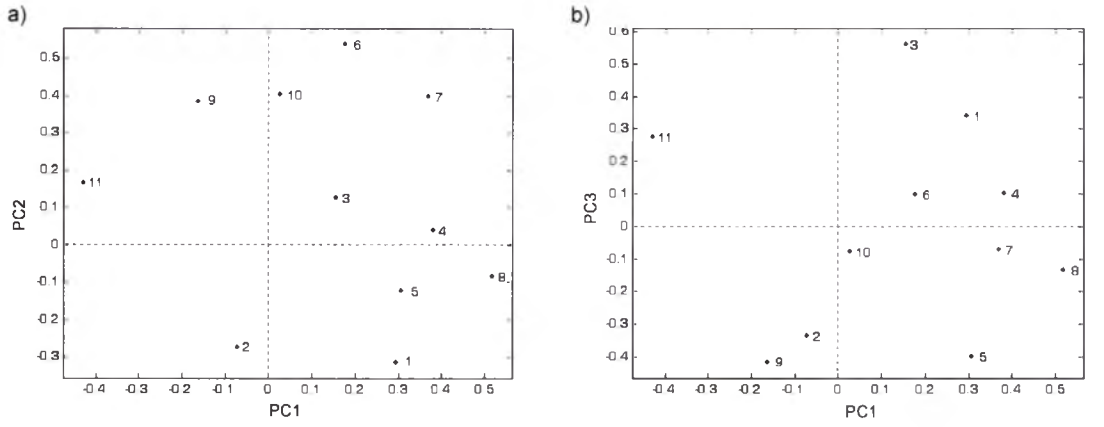
Obiekt odległy został usunięty z danych, a dane poddano ponownej analizie czynników głównych. Nie stwierdzono występowania innych obiektów odległych ani zależności pomiędzy jakością wina, a wartościami czynników głównych. Wyniki analizy widoczne są na poniższych rysunkach.



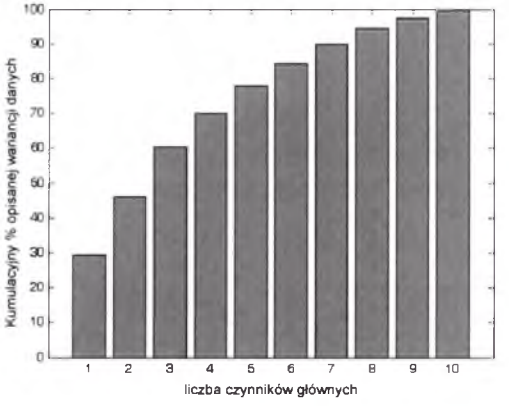
Rys. 75 Projekcja 337 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 76 Projekcja 337 obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono jakość wina (klasa 1 – niska jakość, klasa 2 – wysoka jakość)



Rys. 77 Projekcja parametrów na płaszczyznę zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny, gdzie: 1 – kwasowość trwała, 2 – kwasowość przemijająca, 3 – zawartość kwasu cytrynowego, 4 – pozostałości cukru, 5 – chlorki, 6 – wolny dwutlenek siarki, 7 – całkowita zawartość tlenku siarki, 8 – gęstość, 9 – pH, 10 – zawartość siarczanów i 11 – zawartość alkoholu



Rys. 78 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

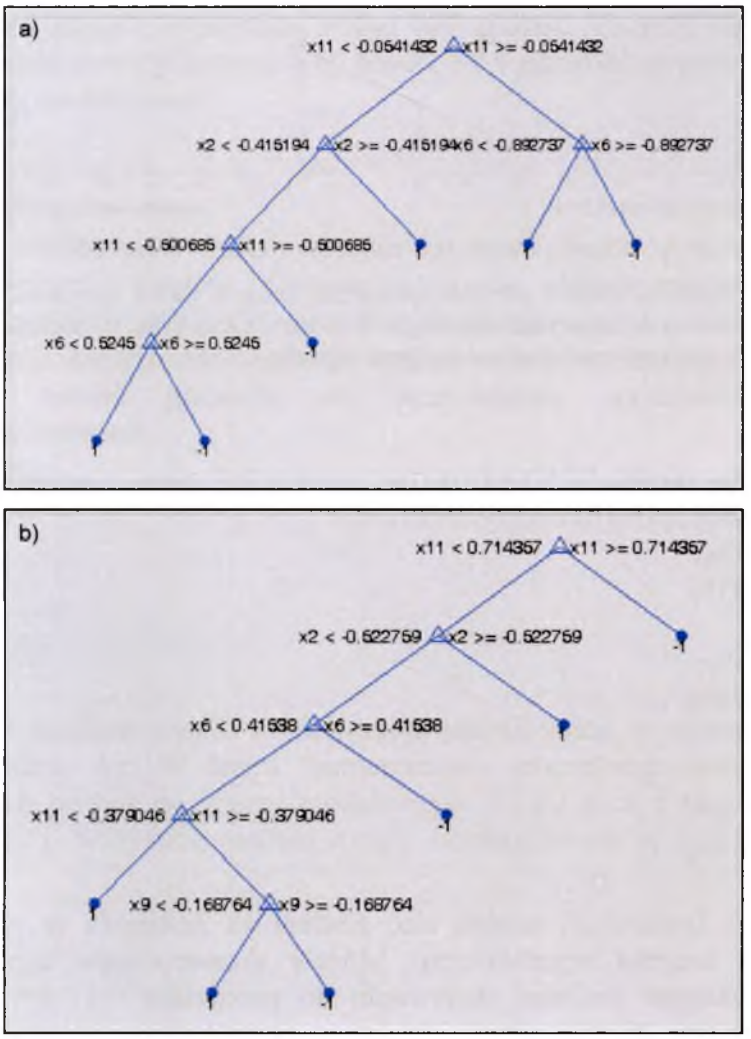
Aby przygotować dane do modelowania podzielono je na trzy zbiory przypisując po 100 obiektów z każdej klasy do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$), po 30 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz resztę (32 obiekty z klasy 1 oraz 45 z klasy 2) do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Tak utworzone zbiory zostały poddane analizie metodą CART oraz PLS.

Drzewa klasyfikacji regresji (A)

Model CART o optymalnej strukturze to drzewo z sześcioma węzłami terminalnymi dla danych zawierających zbiory tworzone za pomocą algorytmu

Kennarda i Stone'a (Rys. 79a). Zmienne wybrane w tym modelu to kwasowość przemijająca (zmienna 2), zawartość wolnego dwutlenku siarki (zmienna 6) i zawartość alkoholu (zmienna 11). Drzewo miało sześć węzłów terminalnych. Procentowe wartości poprawnie sklasyfikowanych próbek przez model wyniosły odpowiednio:
$CCR_{(KS)} = 84{,}85\%$;
$CCRT_{(KS)} = 92{,}31\%$.

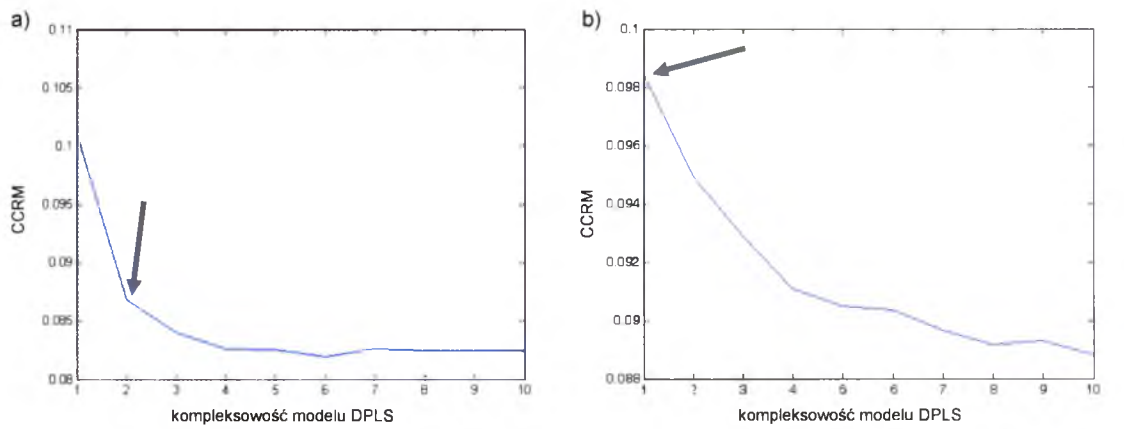
Dla danych zawierających zbiory utworzone za pomocą algorytmu Duplex model CART miał sześć węzłów terminalnych (Rys. 79b), a wskazane zmienne to kwasowość przemijająca (zmienna 2), zawartość wolnego dwutlenku siarki (zmienna 6), pH (zmienna 9) i zawartość alkoholu (zmienna 11). Procentowe wartości poprawnie sklasyfikowanych próbek przez model wyniosły:
$CCR_{(DU)} = 69{,}70\%$;
$CCRT_{(DU)} = 80{,}49\%$.



Rys. 79 Optymalne drzewo CART skonstruowane celem rozróżnienia win pod względem jakości w oparciu o zbiory utworzone za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie (1) klasa 1 i (-1) klasa 2

Metoda częściowych najmniejszych kwadratów w wariancie dyskryminacyjnym (A)

Do konstrukcji modelu DPLS dla danych zawierających zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 80a) wybrano dwa czynniki ukryte w oparciu o zbiór monitoringowy. Kompleksowość modelu dla danych zawierających zbiory otrzymane algorytmem Duplex wyniosła jeden czynnik (DU, Rys. 80b). Przy określaniu kompleksowości modeli wzięto pod uwagę niewielkie różnice błędu CCRM mające miejsce dopiero na trzecim miejscu po przecinku.



Rys. 80 Wykres zależności CCRM od kompleksowości modelu DPLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Ostateczny model DPLS charakteryzowany był przez następujące procentowe wartości poprawnie sklasyfikowanych próbek:
$CCR_{(KS)} = 86{,}00\%$;
$CCRT_{(KS)} = 82{,}05\%$
oraz
$CCR_{(DU)} = 84{,}65\%$;
$CCRT_{(DU)} = 90{,}24\%$.

Sieci neuronowe (A)

Użyta do konstrukcji modeli sieć neuronowa zawierała w węzłach obydwu warstw funkcję tangens hiperboliczny. Modele skonstruowane zostały w oparciu o oryginalne zmienne poddane skalowaniu do przedziału <-1, 1>. Jako pierwszy modelowany zestaw danych użyto obiektów podzielonych na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalna sieć zawierała jedenaście węzłów wejściowych i po jednym węźle w warstwie ukrytej oraz wyjściowej. Sieć ta pozwoliła na rozróżnienie próbek wina pod względem jakości z następującym sukcesem:
$CCR_{(KS)} = 92{,}00\%$;
$CCRT_{(KS)} = 93{,}59\%$.

Drugi zestaw danych zawierał obiekty przydzielone do zbiorów przez algorytm Duplex. Optymalny model skonstruowany dla tych danych zawierał jedenaście węzłów wejściowych i po jednym węźle w warstwie ukrytej i wyjściowej. Sieć ta pozwoliła na przewidzenie modelowanej własności z następującym powodzeniem:
$CCR_{(DU)} = 90{,}10\%$;
$CCRT_{(DU)} = 95{,}12\%$.

Neuronowe systemy rozmyte (A)

Skonstruowano modele NFS typu Sugeno pierwszego rzędu. Jako pierwszy skonstruowano model dla danych zawierających zbiory utworzone za pomocą algorytmu Kennarda i Stone'a. Optymalny model wykorzystywał metodę grupowania różnicowego (o promieniu 0,4) do podziału przestrzeni danych. W ramach tego modelu skonstruowano 141 reguł logicznych, co nie jest zbyt dużą liczbą, jeśli brać pod uwagę liczebność zbioru danych wynoszącą ponad 300 próbek. Uczenie modelu odbywało się z zastosowaniem metody hybrydowej. Model NFS pozwolił na przewidzenie jakości wina z następującym sukcesem:
$CCR_{(KS)} = 100\%$;
$CCRT_{(KS)} = 96{,}15\%$.

Kolejny modelowany zestaw danych zawierał obiekty podzielone na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano cztery reguły logiczne, a uczenie modelu odbywało się według metody wstecznej propagacji błędu. Skonstruowany model pozwolił na przewidzenie modelowanej własności z następującym sukcesem:
$CCR_{(DU)} = 89{,}60\%$;
$CCRT_{(DU)} = 92{,}68\%$.

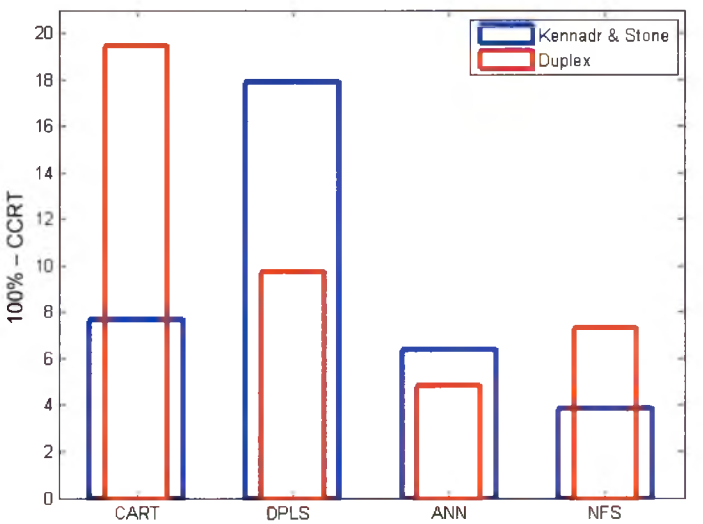
Podsumowanie (A)

Tabela 7 zawiera wyniki modelowania jakości wina w oparciu o jego skład chemiczny (Zestaw A). W tabeli zamieszczono procentowe wartości poprawnie sklasyfikowanych próbek ze zbioru modelowego (CCR) oraz z niezależnego zbioru testowego (CCRT). Wszystkie modele zostały skonstruowane w oparciu o oryginalne zmienne.

Tabela 7 Zestawienie wyników przeprowadzonych analiz dla modelowania jakości wina (Dane 6 A), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	CCR [%]	CCRT [%]	opis modelu
CART	KS	oryginalne	84,85	92,31	6 węzłów terminalnych
	DU	oryginalne	69,70	80,49	6 węzłów terminalnych
DPLS	KS	oryginalne	86,00	82,05	2 czynniki ukryte
	DU	oryginalne	84,65	90,24	1 czynnik ukryty
ANN	KS	oryginalne	92,00	93,39	1 węzeł w warstwie ukrytej
	DU	oryginalne	90,10	95,12	1 węzeł w warstwie ukrytej
NFS	KS	oryginalne	100	96,15	141 reguł logicznych
	DU	oryginalne	89,60	92,68	4 reguły logiczne

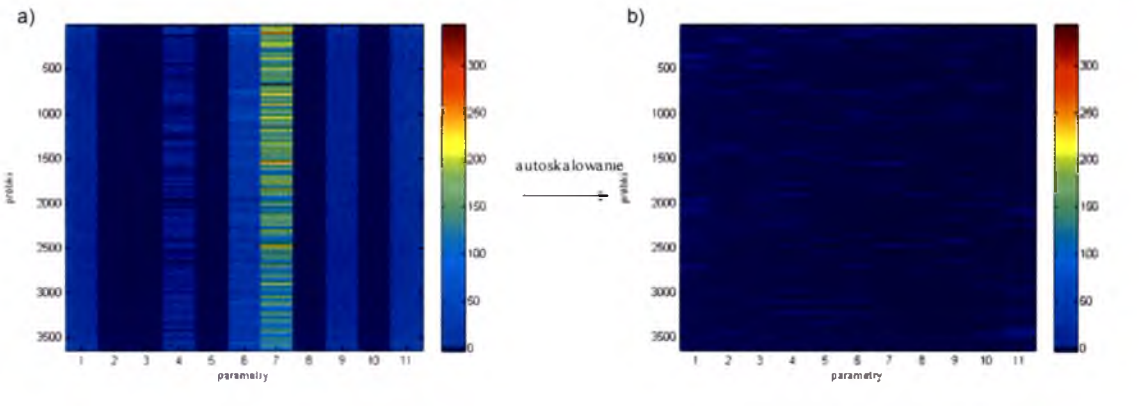
Na poniższym wykresie (Rys. 82) widoczne jest, iż wyniki dla modelu NFS były lepsze od wyników dla metody CART i DPLS niezależnie od sposobu podziału obiektów na zbiory (KS czy DU). Metoda NFS pozwoliła także na konstrukcję modeli o porównywalnej mocy predykcyjnej z metodą ANN.



Rys. 81 Wykres procentu błędnie sklasyfikowanych próbek (100% – CCR) charakteryzujący konstruowane modele

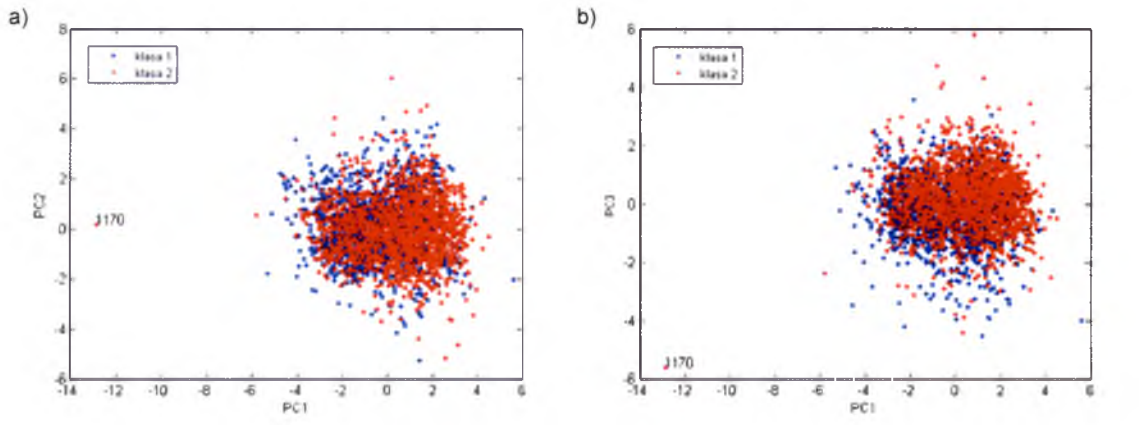
Zestaw B

Wymiarowość danych z zestawu B to 3655 x 11. Zmienna zależna **y** kodująca przynależność obiektów do klas miała postać binarną. Dane wstępnie przygotowano poddając je operacji autoskalowania (Rys. 82).



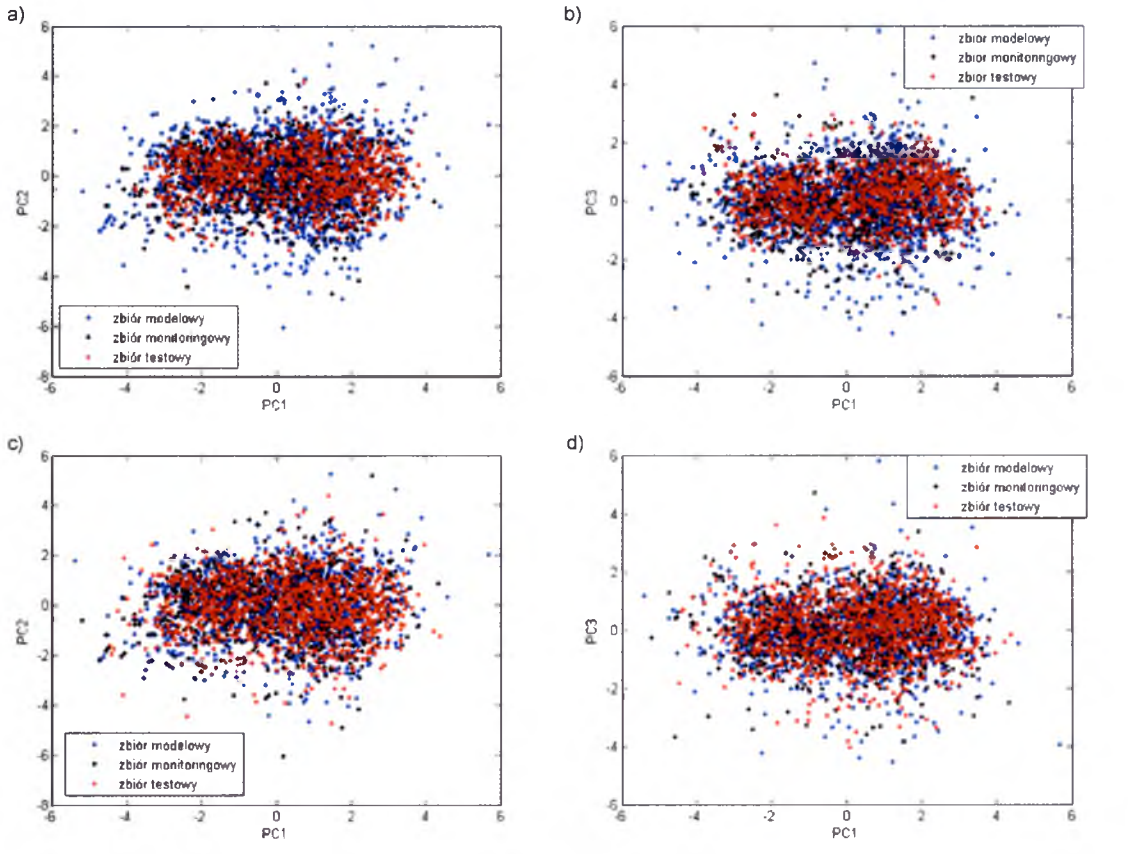
Rys. 82 Wartości jedenastu parametrów dla 3655 próbek wina a) przed i b) po autoskalowaniu

Eksploracja i przygotowanie danych (B)

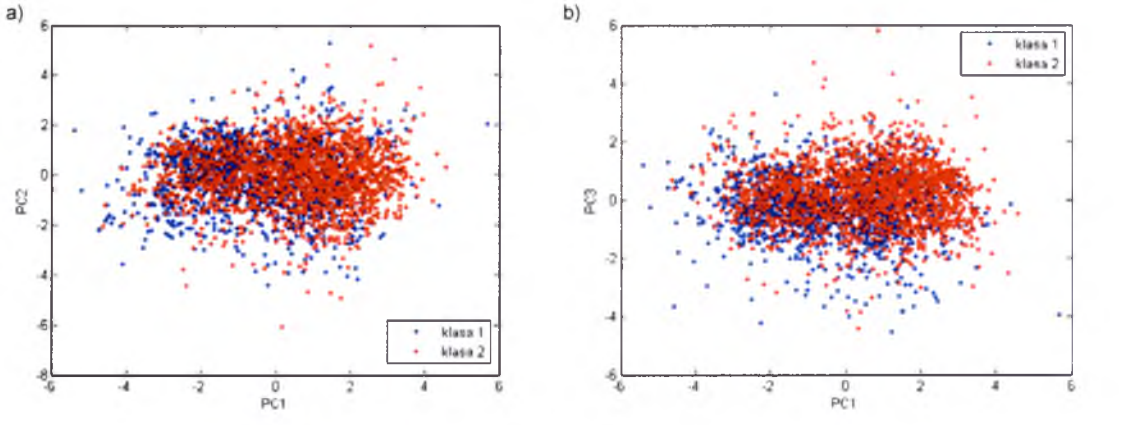


Rys. 83 Projekcja 3655 obiektów na płaszczyznę zdefiniowaną przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny; gdzie zaznaczono obiekt odległy nr 1170

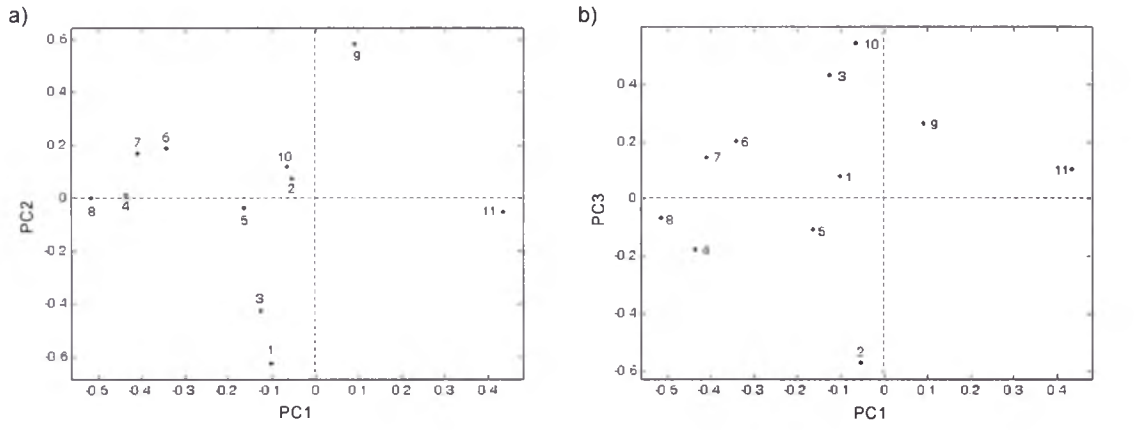
Analiza czynników głównych pozwoliła na wykrycie obiektu odległego w danych. Była to próbka nr 1170 (Rys. 83). Po usunięciu próbki dane ponownie zanalizowano wykorzystując do tego celu metodę PCA. Innych obiektów odległych nie stwierdzono (Rys. 84-87). Rys. 86 ukazuje skorelowane parametry, są to następujące pary zmiennych: 6 i 7, 4 i 8 oraz 2 i 10. Wymiarowość danych poddanych modelowaniu wynosiła 3654 x 11, z czego 1456 próbek należało do grupy pierwszej. Liczba próbek w drugiej grupie nie uległa zmianie (2198 obiektów).



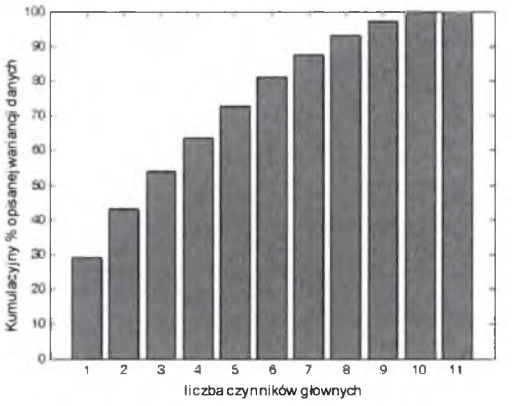
Rys. 84 Projekcja 3654 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 85 Projekcja 3654 obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono jakość wina (klasa 1 – niższa jakość, klasa 2 – wyższa jakość)



Rys. 86 Projekcja parametrów na płaszczyznę zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny, gdzie: 1 – kwasowość trwała, 2 – kwasowość przemijająca, 3 – zawartość kwasu cytrynowego, 4 – pozostałości cukru, 5 – chlorki, 6 – wolny dwutlenek siarki, 7 – całkowita zawartość tlenku siarki, 8 – gęstość, 9 – pH, 10 – zawartość siarczanów i 11 – zawartość alkoholu



Rys. 87 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

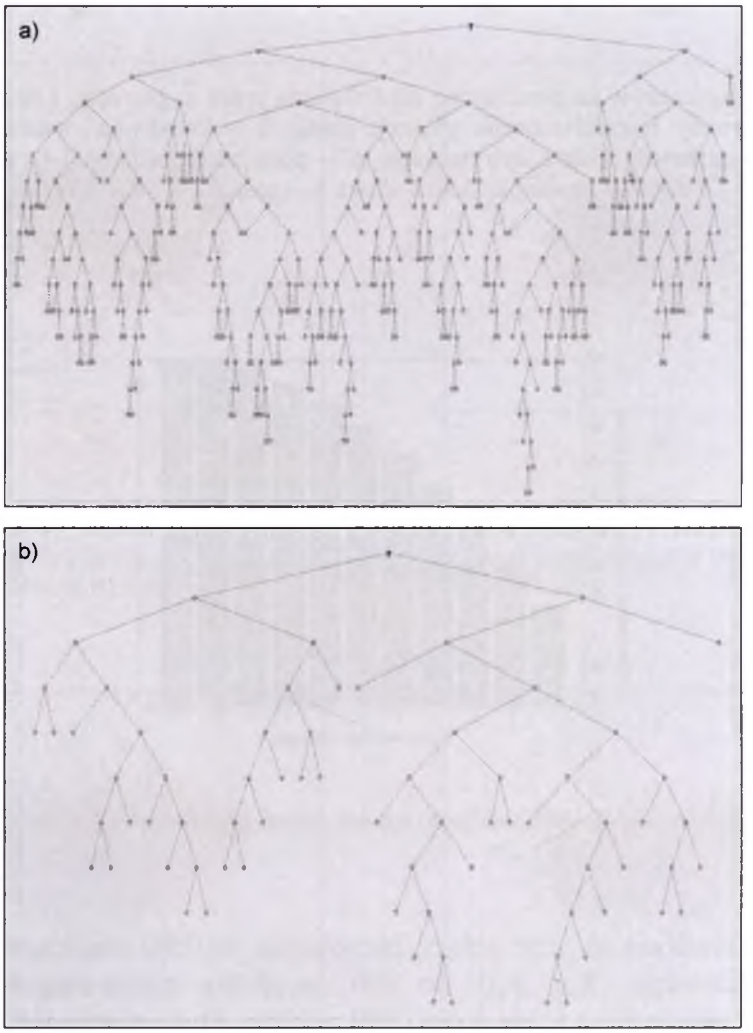
Dane podzielono na trzy zbiory przypisując po 900 obiektów z każdej klasy do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$), po 300 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz resztę (257 obiektów z klasy 1 oraz 991 z klasy 2) do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Zbiory utworzono za pomocą algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Następnie dane poddano procesowi modelowania.

Drzewa klasyfikacji regresji (B)

Najlepszy model CART to drzewo z 202 węzłami terminalnymi (Rys. 88a) dla danych zawierających zbiory utworzone za pomocą algorytmu Kennarda i Stone'a. Jako zmienne decyzyjne przez model zostały wskazane wszystkie zmienne. Procentowe wartości poprawnie sklasyfikowanych próbek przez model wyniosły odpowiednio:

$CCR_{(KS)} = 79{,}38\%$;
$CCRT_{(KS)} = 65{,}23\%$.

Model CART dla danych zawierających zbiory utworzone za pomocą algorytmu Duplex miał trzydzieści siedem węzłów terminalnych (Rys. 88b). Do wskazanych zmiennych decyzyjnych nie należała tylko zmienna nr 1 – kwasowość trwała. Procentowe wartości poprawnie sklasyfikowanych próbek przez model wyniosły:
$CCR_{(DU)} = 60{,}70\%$;
$CCRT_{(DU)} = 70{,}49\%$.

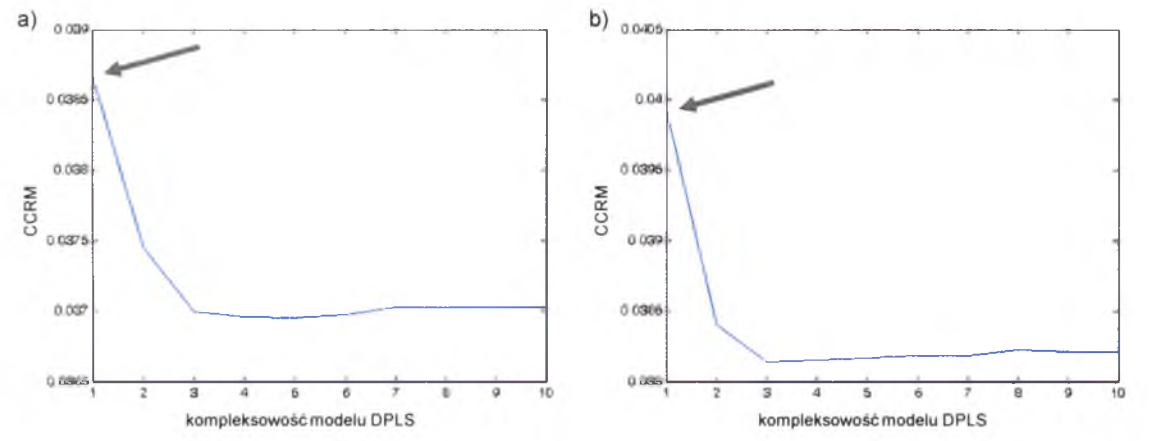


Rys. 88 Optymalne drzewo CART skonstruowane celem klasyfikacji win pod względem jakości w oparciu o zbiory utworzone za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU)

Metoda częściowych najmniejszych kwadratów w wariancie dyskryminacyjnym (B)

Dyskryminacyjne modele PLS skonstruowano dla danych zawierających zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 89a) i algorytmu Duplex (DU, Rys. 89b). Wybór kompleksowości modelu odbywa się w oparciu o CCRM,

którego wartości dla kolejnych czynników głównych różnią się nieznacznie od siebie w tym konkretnym przypadku. Dlatego też w obydwu przypadkach skonstruowano modele z jednym czynnikiem ukrytym. Modele te charakteryzowane były przez następujące procentowe wartości poprawnie sklasyfikowanych próbek:
$CCR_{(KS)} = 62,67\%$;
$CCRT_{(KS)} = 67,22\%$
oraz
$CCR_{(DU)} = 65,22\%$;
$CCRT_{(DU)} = 62,44\%$.



Rys. 89 Wykres zależności CCRM od kompleksowości modelu DPLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Sieci neuronowe (B)

Konstruowane sieci neuronowe miały jedenaście węzłów wejściowych, dwa węzły w warstwie ukrytej i jeden węzeł w warstwie wyjściowej. Węzły warstwy ukrytej oraz wyjściowej były wyposażone w funkcję tangens hiperboliczny. Modele konstruowane były w oparciu o oryginalne zmienne poddane skalowaniu do przedziału <-1, 1>. W pierwszej kolejności modelowaniu poddano zestaw danych utworzony z obiektów podzielonych na zbiory za pomocą algorytmu Kennarda i Stone'a, potem algorytmu Duplex. Optymalna struktura modeli ANN to w obydwu przypadkach jedenaście węzłów wejściowych, dwa węzły w warstwie ukrytej oraz jeden węzeł wyjściowy. Opracowane sieci ANN pozwoliły na przewidzenie jakości wina z następującym sukcesem:
$CCR_{(KS)} = 51,50\%$;
$CCRT_{(KS)} = 79,98\%$.
oraz
$CCR_{(DU)} = 51,56\%$;
$CCRT_{(DU)} = 79,58\%$.

Neuronowe systemy rozmyte (B)

Jako pierwszy skonstruowano model NFS typu Sugeno pierwszego rzędu dla danych zawierających zbiory utworzone za pomocą algorytmu Kennarda i Stone'a. Optymalny model wykorzystywał metodę grupowania różnicowego (o promieniu 0,2) do podziału przestrzeni danych. W ramach tego modelu skonstruowano 29 reguł logicznych. Uczenie modelu odbywało się z zastosowaniem metody wstecznej propagacji błędu. Skonstruowany model pozwolił na przewidzenie jakości wina z następującym sukcesem:
$CCR_{(KS)} = 79{,}11\%$;
$CCRT_{(KS)} = 75{,}52\%$.

Drugi modelowany zestaw danych zawierał obiekty podzielone na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano cztery reguły logiczne, a uczenie modelu odbywało się według metody hybrydowej. Skonstruowany model NFS pozwolił na przewidzenie modelowanej własności z następującym powodzeniem:
$CCR_{(DU)} = 66{,}33\%$;
$CCRT_{(DU)} = 71{,}77\%$.

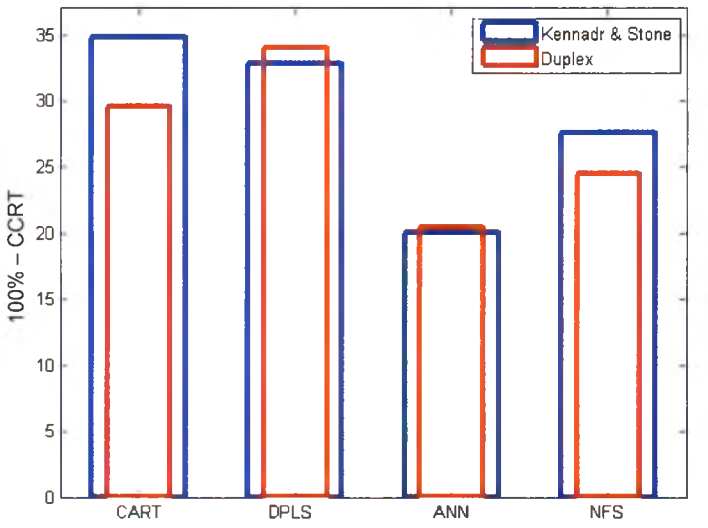
Podsumowanie (B)

Wyniki modelowania jakości wina w oparciu o jego skład chemiczny dla zestawu B przedstawiono w poniższej tabeli. Procentowe wartości poprawnie sklasyfikowanych próbek ze zbioru modelowego (CCR) oraz z niezależnego zbioru testowego (CCRT) zamieszczono odpowiednio w czwartej i piątej kolumnie. Wszystkie modele zostały skonstruowane w oparciu o oryginalne zmienne.

Tabela 8 Zestawienie wyników przeprowadzonych analiz dla modelowania jakości wina (Dane 6 B), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	CCR [%]	CCRT [%]	opis modelu
CART	KS	oryginalne	79,38	65,23	202 węzły terminalne
	DU	oryginalne	60,70	70,49	37 węzłów terminalnych
DPLS	KS	oryginalne	62,67	67,22	1 czynnik ukryty
	DU	oryginalne	65,22	62,44	1 czynnik ukryty
ANN	KS	oryginalne	51,50	79,98	2 węzły w warstwie ukrytej
	DU	oryginalne	51,56	79,58	2 węzły w warstwie ukrytej
NFS	KS	oryginalne	79,11	75,22	29 reguł logicznych
	DU	oryginalne	66,33	71,77	4 reguły logiczne

Rys. 90 przedstawia procentowe ilości błędnie sklasyfikowanych próbek z niezależnego zbioru testowego. Metoda NFS pozwolił na konstrukcję modeli o lepszej mocy predykcyjnej niż metoda CART i DPLS. Z drugiej jednak strony modele NFS obarczone były nieco większym błędem niż modele ANN.

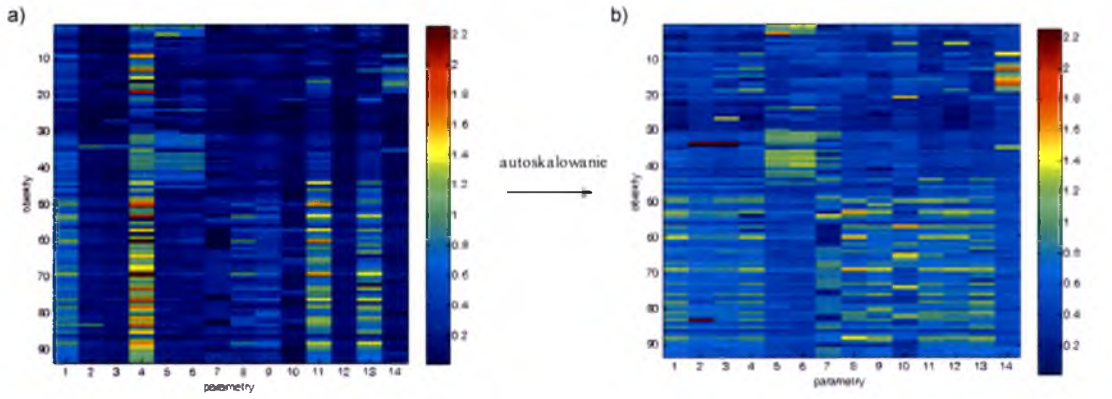


Rys. 90 Wykres procentu błędnie sklasyfikowanych próbek (100% – CCR) charakteryzujący konstruowane modele

9.7 Dane 7: Modelowanie pochodzenia opium

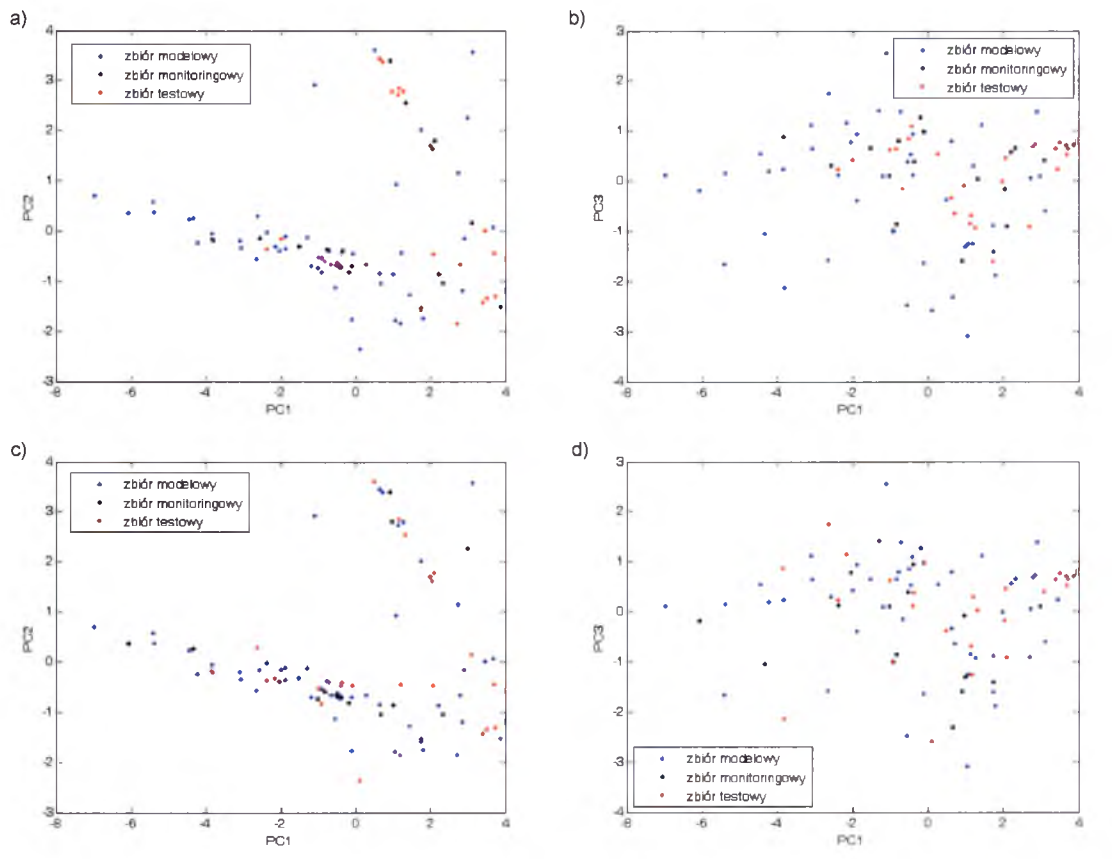
Oznaczono zawartość 14 aminokwasów w 93 próbkach indyjskiego opium za pomocą chromatografii cieczowej sprzężonej z detektorem fluorometrycznym [111]. Próbki pochodziły z dwóch regionów Indii: 51 próbek z regionu Rajasthan i 42 z regionu Madhya Prasesh. Do oznaczonych aminokwasów należały: kwas asparaginowy (D), treonina (T), seryna (S), kwas glutaminowy (E), glicyna (G), alanina (A), walina (V), izoleucyna (I), leucyna (L), tyrozyna (Y), fenyloalanina (F), histydyna (H), lizyna (K) i arginina (R).

Wymiarowość analizowanych danych to 93 **x** 14. Pochodzenie geograficzne zakodowano w postaci binarnej zmiennej zależnej **y**. Dane poddano autoskalowaniu (Rys. 91).

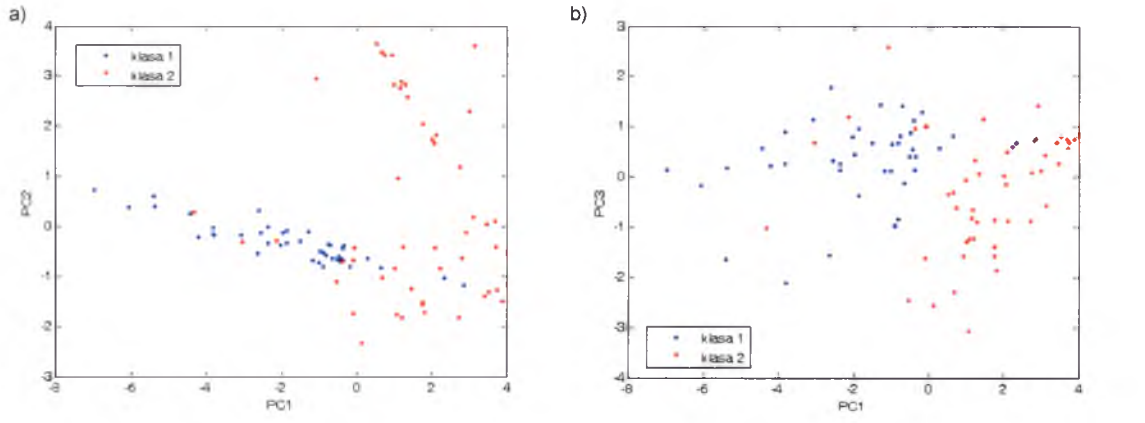


Rys. 91 Wartości czternastu parametrów dla 93 próbek opium a) przed i b) po autoskalowaniu, gdzie: 1 – kwas asparaginowy, 2 – treonina, 3 – seryna, 4 – kwas glutaminowy, 5 – glicyna, 6 – alanina, 7 – walina, 8 – izoleucyna, 9 – leucyna, 10 – tyrozyna, 11 – fenyloalanina, 12 – histydyna, 13 – lizyna i 14 – arginina

Eksploracja i przygotowanie danych

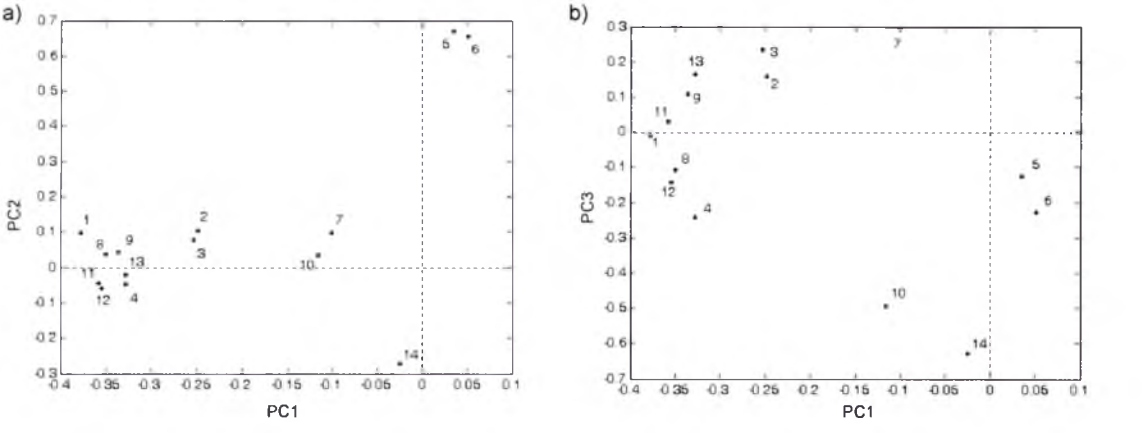


Rys. 92 Projekcja 93 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex

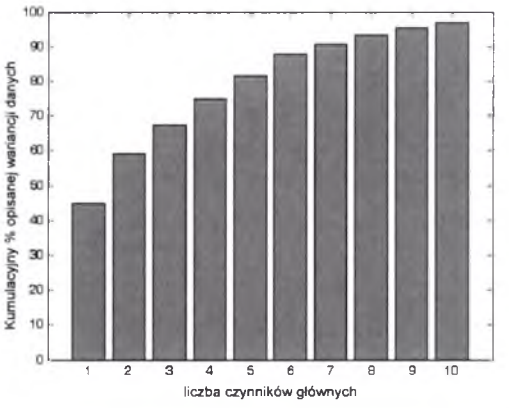


Rys. 93 Projekcja 93 obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono pochodzenie opium (klasa 1 – Rajasthan, klasa 2 – Madhya Prasesh)

Metoda PCA pozwoliła na wykluczenie występowania obiektów odległych w przestrzeni mierzonych parametrów oraz innych niepożądanych efektów (Rys. 92-95). Rozmieszczenie obiektów należących do poszczególnych klas w przestrzeni czynników głównych można częściowo powiązać z wartościami pierwszego czynnika głównego (PC1). Na rysunku 93 widoczne jest, iż większość obiektów należących do pierwszej klasy ma ujemne wartości na osi PC1, natomiast obiekty z klasy drugiej mają dodatnie wartości. Z kolei projekcja parametrów na płaszczyznę zdefiniowaną przez pierwsze dwa czynniki główne (Rys. 94a) pokazuje, dwie grupy skorelowanych parametrów, które są ortogonalne względem siebie. Do pierwszej grupy parametrów należą skorelowane dodatnio zmienne 5 i 6 oraz skorelowana do nich ujemnie zmienna 14. Grupę drugą stanowią pozostałe parametry. Prawidłowość ta, choć mniej wyraźna zauważalna, jest także na rysunku 94b.



Rys. 94 Projekcja parametrów na płaszczyznę wag zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny, gdzie: 1 – kwas asparaginowy, 2 – treonina, 3 – seryna, 4 – kwas glutaminowy, 5 – glicyna, 6 – alanina, 7 – walina, 8 – izoleucyna, 9 – leucyna, 10 – tyrozyna, 11 – fenyloalanina, 12 – histydyna, 13 – lizyna i 14 – arginina

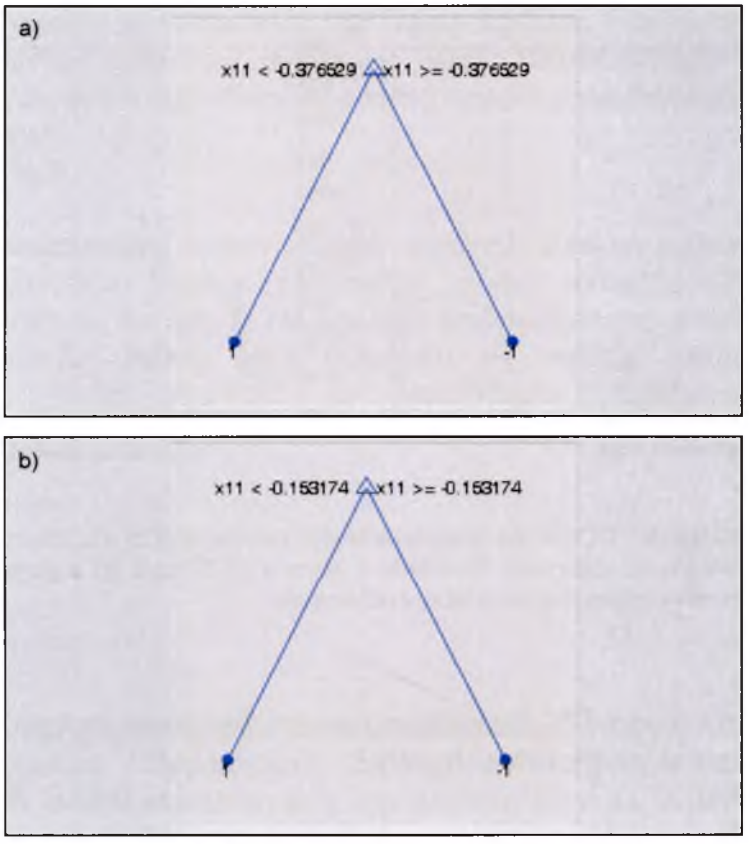


Rys. 95 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

Przed modelowaniem dane poddano wstępnemu przygotowaniu do analizy. W tym celu obiekty podzielono na trzy zbiory przypisując po 25 obiektów z każdej klasy do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$), po 9 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz resztę (17 obiektów z klasy 1 oraz 8 z klasy 2) do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Następnie tak przygotowane dane zostały poddane analizie metodą CART oraz PLS.

Drzewa klasyfikacji regresji

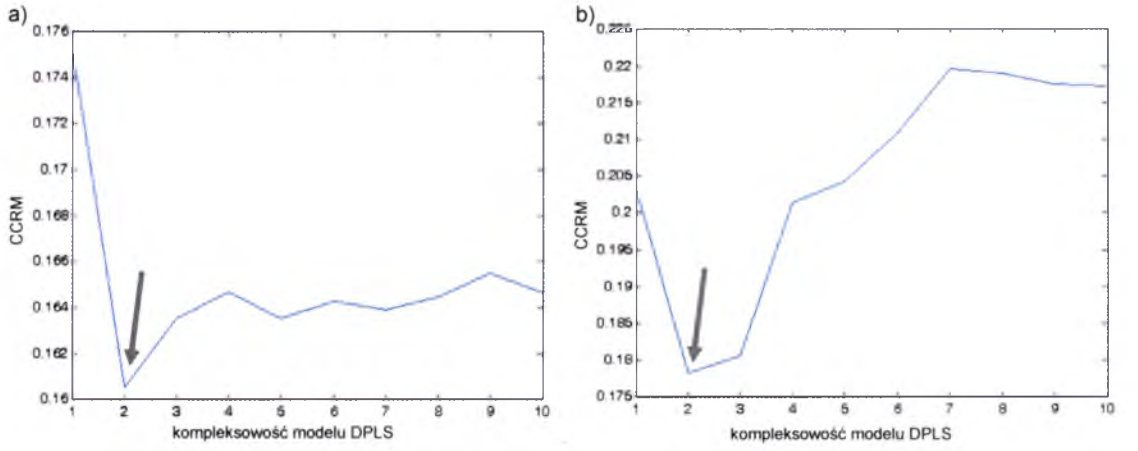
Optymalne struktury drzew CART (Rys. 96) miały po dwa węzły terminalne zarówno dla modeli konstruowanych w oparciu o dane zawierające zbiory utworzone za pomocą algorytmu Kennarda i Stone'a jak i algorytmu Duplex. Obydwa modele wskazały zawartość fenyloalaniny (zmienna 11) jako parametr decyzyjny. Procentowe wartości poprawnie sklasyfikowanych próbek przez model wynosiły odpowiednio:
$CCR_{(KS)} = 100\%$;
$CCRT_{(KS)} = 100\%$
oraz
$CCR_{(DU)} = 94{,}12\%$;
$CCRT_{(DU)} = 96{,}00\%$.



Rys. 96 Optymalne drzewo CART skonstruowane celem oznaczania pochodzenia opium w oparciu o zbiory utworzone za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie (1) klasa 1 i (-1) klasa 2

Metoda częściowych najmniejszych kwadratów w wariancie dyskryminacyjnym

Kompleksowość modelu DPLS wynosiła dwa czynniki ukryte dla danych zawierających zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 97a). Dla modelu konstruowanego w oparciu o dane zawierające zbiory otrzymane algorytmem Duplex wybrane zostały także dwa czynniki ukryte (DU, Rys. 97b).



Rys. 97 Wykres zależności CCRM od kompleksowości modelu DPLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Optymalny model DPLS charakteryzowany był przez następujące procentowe wartości poprawnie sklasyfikowanych próbek:
$CCR_{(KS)} = 84{,}00\%$;
$CCRT_{(KS)} = 100\%$
oraz
$CCR_{(DU)} = 92{,}00\%$;
$CCRT_{(DU)} = 92{,}00\%$.

Sieci neuronowe

Niezależnie od sposobu podziału obiektów na zbiory skonstruowane modele sieci neuronowych wykorzystywały funkcję tangens hiperboliczny jako funkcję aktywacji węzłów zarówno warstwy ukrytej, jak i wyjściowej. Każda z sieci miała czternaście węzłów wejściowych, po jednym ukrytym oraz jednym wyjściowym. Modele konstruowane były w oparciu o oryginalne zmienne poddane skalowaniu do przedziału od -1 do 1. Skonstruowane sieci pozwoliły na określenie pochodzenia próbek opium z następującym sukcesem:
$CCR_{(KS)} = 77{,}00\%$;
$CCRT_{(KS)} = 100\%$;
oraz
$CCR_{(DU)} = 94{,}00\%$;
$CCRT_{(DU)} = 100\%$.

Neuronowe systemy rozmyte

Jako pierwszy skonstruowano model NFS typu Sugeno pierwszego rzędu, dla danych zawierających zbiory utworzone za pomocą algorytmu Kennarda i Stone'a. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych.

W ramach tego modelu skonstruowano trzy reguły logiczne. Uczenie modelu odbywało się z zastosowaniem wstecznej propagacji błędu. Skonstruowany model pozwolił na przewidzenie modelowanej własności z następującym powodzeniem:
$CCR_{(KS)} = 84{,}00\%$;
$CCRT_{(KS)} = 100\%$.

Drugi modelowany zestaw danych zawierał obiekty podzielone na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystuje metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano piętnaście reguł logicznych. Uczenie modelu NFS odbywało się według metody hybrydowej. Skonstruowany model pozwolił na przewidzenie modelowanej własności z następującym sukcesem:
$CCR_{(DU)} = 100\%$;
$CCRT_{(DU)} = 96{,}66\%$.

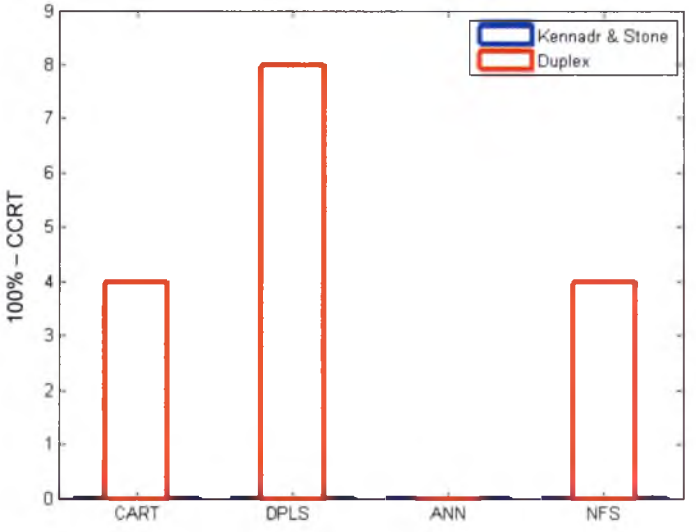
Podsumowanie

Tabela 9 przedstawia wyniki modelowania pochodzenia próbek opium z dwóch regionów w Indiach. Dopasowanie modelu do danych i moc predykcyjną skonstruowanych modeli charakteryzują odpowiednio wartości CCR i CCRT. Modele CART, DPLS, ANN i NFS skonstruowano w oparciu o oryginalne zmienne.

Tabela 9 Zestawienie wyników przeprowadzonych analiz dla modelowania pochodzenia próbek opium (Dane 7), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	CCR [%]	CCRT [%]	opis modelu
CART	KS	oryginalne	100	100	2 węzły terminalne
	DU	oryginalne	94,12	96,00	2 węzły terminalne
DPLS	KS	oryginalne	84,00	100	2 czynniki ukryte
	DU	oryginalne	92,00	92,00	2 czynniki ukryte
ANN	KS	oryginalne	77,00	100	1 węzeł w warstwie ukrytej
	DU	oryginalne	94,00	100	1 węzeł w warstwie ukrytej
NFS	KS	oryginalne	84,00	100	3 reguły logiczne
	DU	oryginalne	100	96,66	15 reguł logicznych

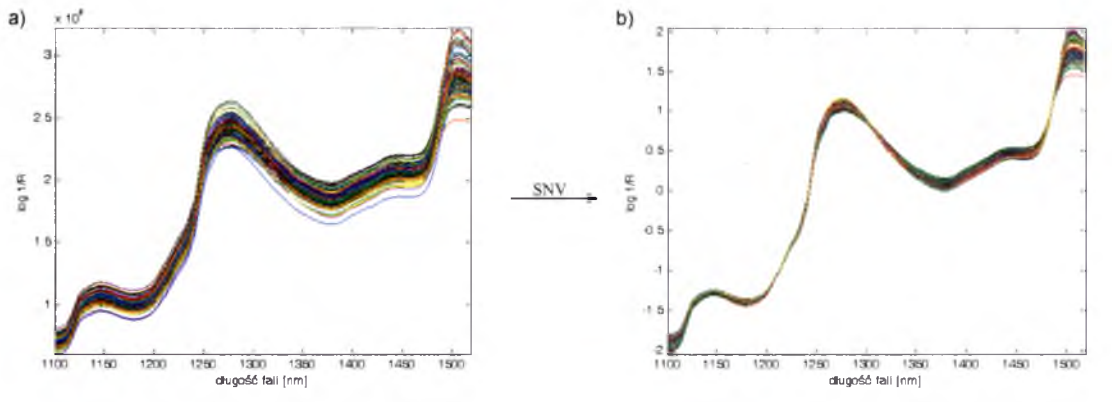
Na rysunku 98 przedstawiono procentowy wykres błędnie sklasyfikowanych próbek za pomocą skonstruowanych modeli dyskryminacyjnych. Wyniki modelowania danych zawierających zbiory utworzone algorytmem Kennarda i Stone'a dla wszystkich metod były takie same – nie stwierdzono błędnie sklasyfikowanych próbek. Z kolei dla danych zawierających zbiory utworzone algorytmem Duplex model NFS był obarczony takim samym błędem jak model CART. Ponadto model NFS przewyższał mocą predykcyjną model DPLS.



Rys. 98 Wykres procentu błędnie sklasyfikowanych próbek (100% – CCR) charakteryzujący konstruowane modele

9.8 Dane 8: Modelowanie składu paszy zwierzęcej

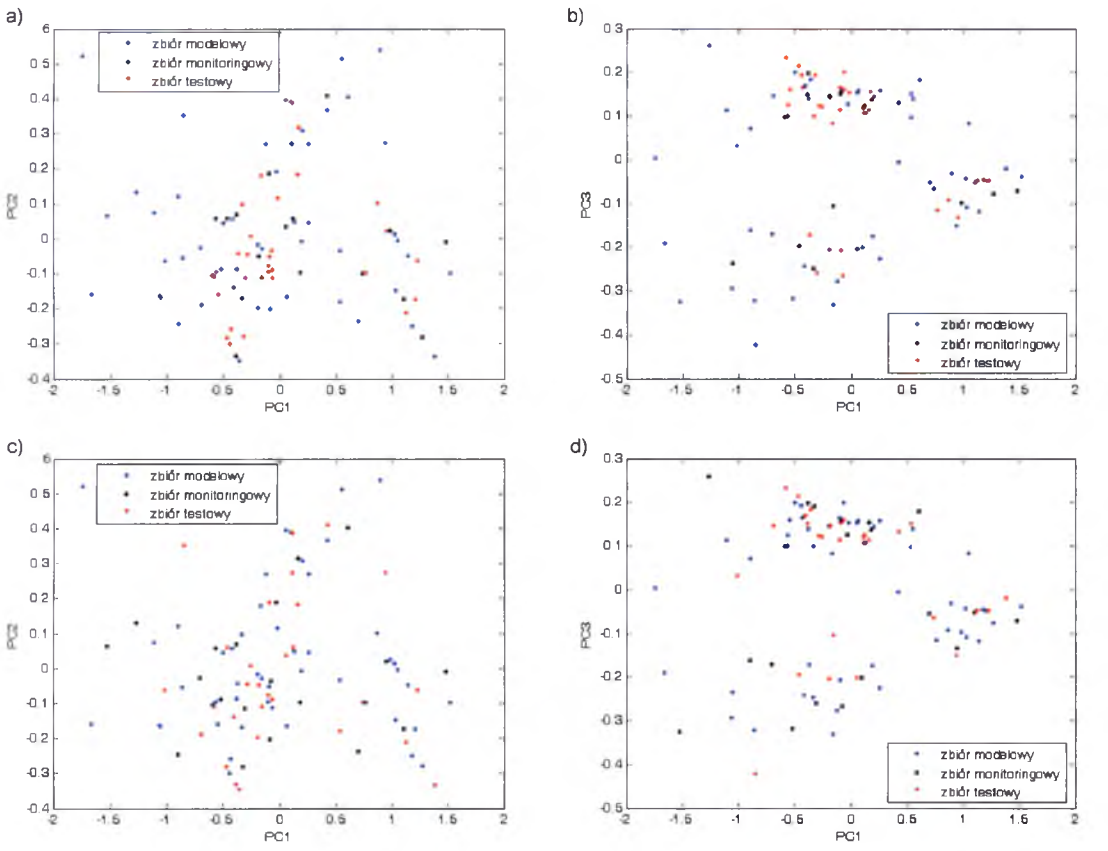
Dane 8 zawierały widma NIR zarejestrowane dla 97 próbek świńskiej wątroby celem oznaczenia rodzaju dodawanego tłuszczu do paszy zwierzęcej [112]. Świnie karmiono paszą zawierającą dodatek w postaci oleju sojowego (53 osobników) lub mieszaniny tłuszczu zwierzęcego i roślinnego (44 sztuki). Widma rejestrowano techniką odbiciowej spektroskopii w bliskiej podczerwieni w zakresie od 1100 nm do 1520 nm. Dane o wymiarowości 97 x 420 poddano transformacji SNV (Rys. 99b).



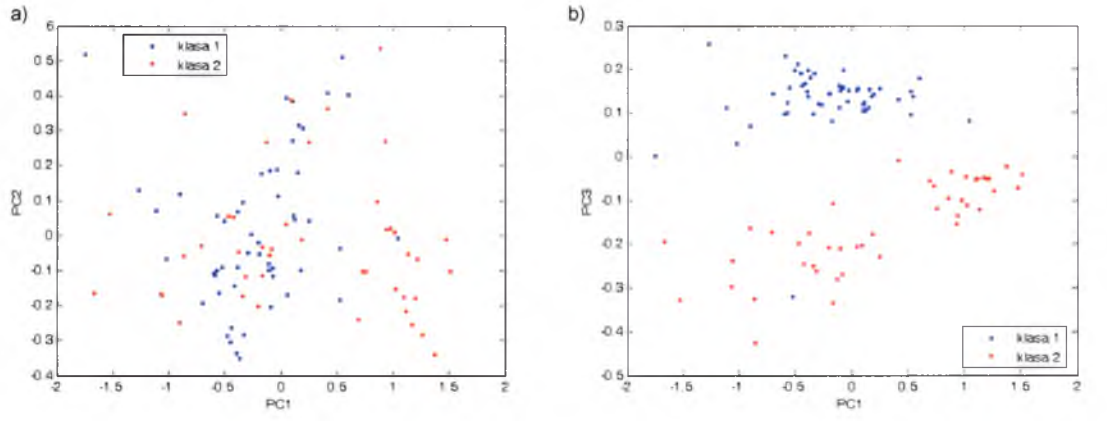
Rys. 99 Widma NIR 97 próbek świńskiej wątroby a) przed i b) po transformacji SNV

Eksploracja i przygotowanie danych

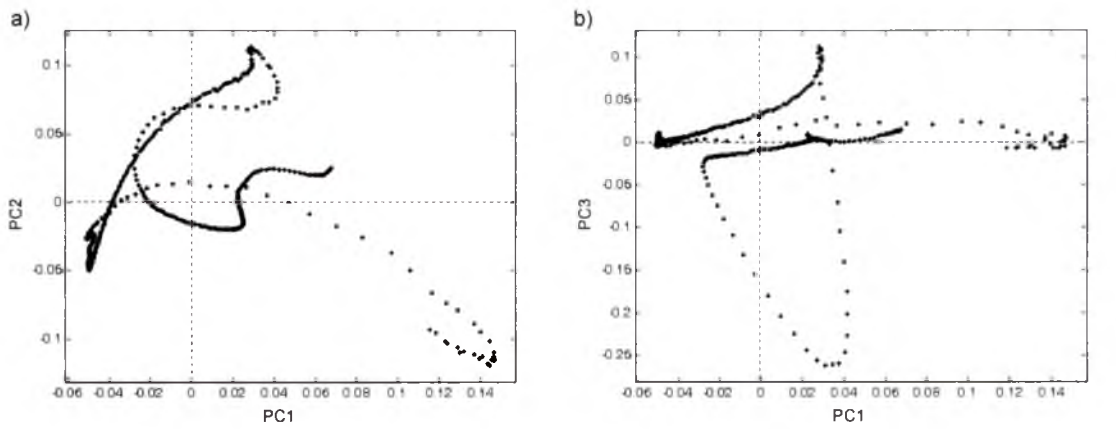
Metoda PCA pozwoliła na wizualizację i eksploracje danych. Rys. 101b pokazuje, że kombinacja pierwszego oraz trzeciego czynnika głównego pozwala na dobre odseparowanie grup obiektów należących do każdej z klas. Pu upewnieniu się, iż nie występują obiekty odległe (Rys. 100-103) przystąpiono do przygotowania danych do modelowania.



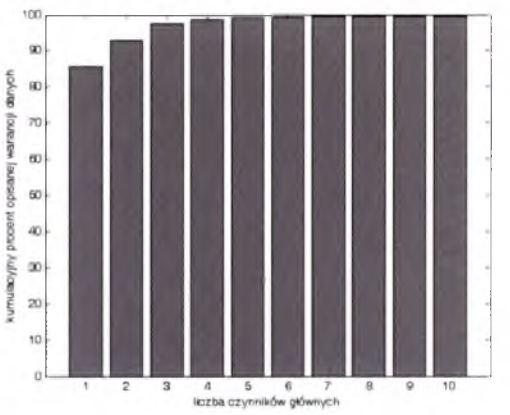
Rys. 100 Projekcja 97 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 101 Projekcja 97 obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono rodzaj dodatku paszowego (klasa 1 – olej sojowy, klasa 2 – mieszanina tłuszczu zwierzęcego i roślinnego)



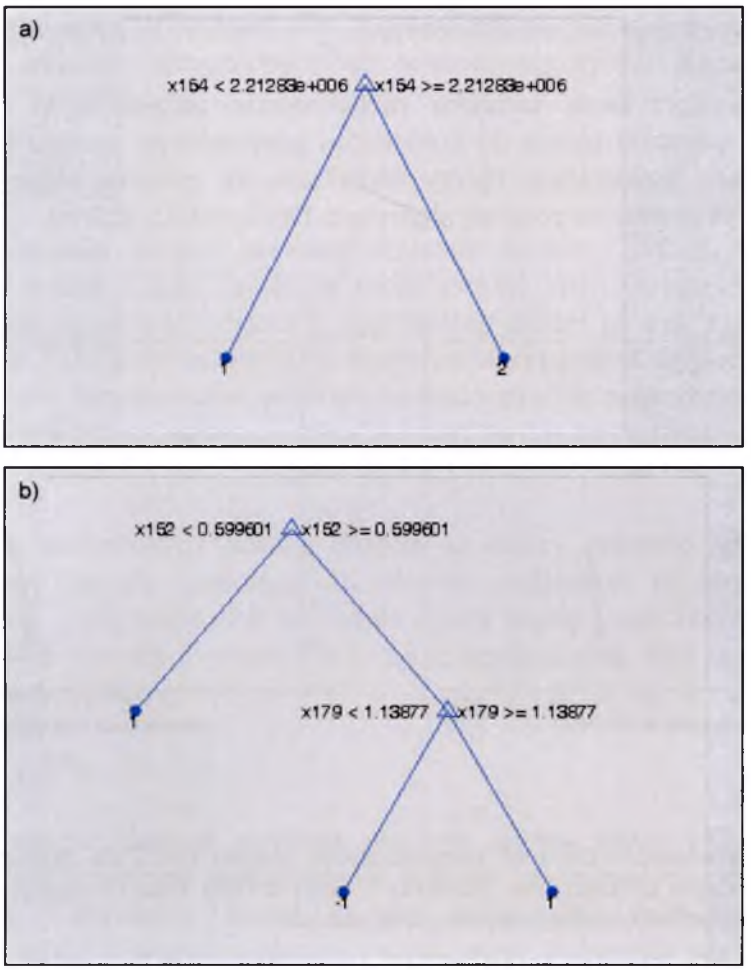
Rys. 102 Projekcja parametrów na płaszczyznę wag zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny



Rys. 103 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

Przygotowanie danych do analizy polegało na podzieleniu obiektów na trzy zbiory (modelowy, monitoringowy i testowy), przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). W zbiorze modelowym ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$) znalazły się po 24 obiekty z każdej klasy, w zbiorze monitoringowym ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) po 10 obiektów, a reszta (19 obiektów z klasy 1 oraz 10 z klasy 2) została przydzielona do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Zmienna zależna dla wszystkich zbiorów ($\mathbf{y}_{ml}$, $\mathbf{y}_{mr}$, $\mathbf{y}_{tt}$) została zakodowana binarnie.

Drzewa klasyfikacji i regresji



Rys. 104 Optymalne drzewo CART skonstruowane celem klasyfikacji dodatków paszowych w oparciu o zbiory utworzone za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie (1) klasa 1 i (-1) klasa 2

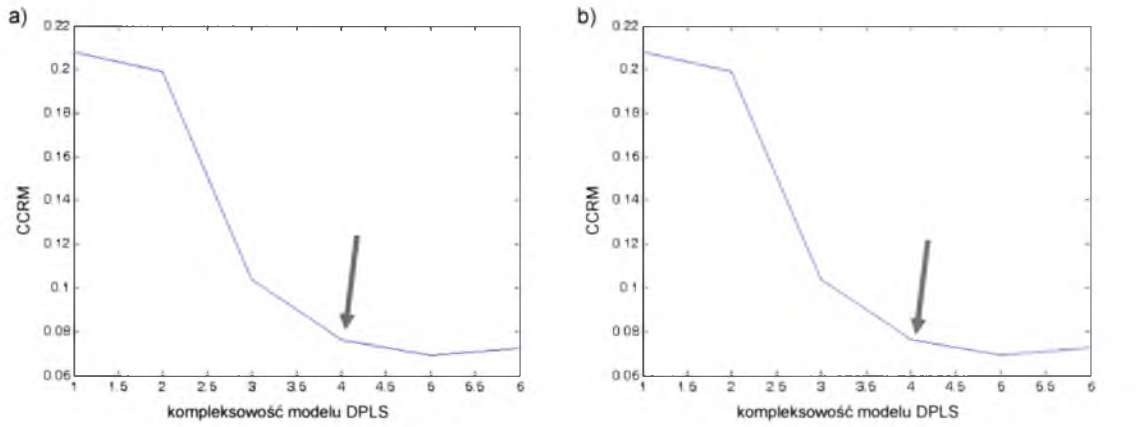
Optymalne binarne drzewo decyzyjne skonstruowane w oparciu o zbiory utworzone za pomocą algorytmu Kennarda i Stone'a miało dwa węzły terminalne (Rys. 104a). Zmienne wskazane w modelu CART jako decyzyjne to zmienna 154 oraz zmienne 1, 151, 174 wskazana przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły odpowiednio:

$CCR_{(KS)} = 74{,}36\%$;
$CCRT_{(KS)} = 100\%$.

Optymalny model skonstruowany w oparciu o dane zawierające zbiory utworzone za pomocą algorytmu Duplex miał trzy węzły terminalne (Rys. 104b). Zmienne wskazane w modelu jako decyzyjne to zmienna 152 i 179. Wartości błędów dla tego modelu wyniosły:
$CCR_{(DU)} = 100\%$;
$CCRT_{(DU)} = 100\%$.

Metoda częściowych najmniejszych kwadratów w wariancie dyskryminacyjnym

Drugą wykorzystaną techniką modelowania danych była metoda DPLS. Wybrano cztery czynniki ukryte do konstrukcji optymalnego modelu konstruowanego w oparciu o dane zawierające zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys a) oraz za pomocą algorytmu Duplex (DU, Rys b).



Rys. 105 Wykres zależności CCRM od kompleksowości modelu DPLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Kompleksowość finalnych modeli DPLS to cztery czynniki ukryte (Rys. 105) bez względu na sposób podziału obiektów na zbiory. Skonstruowane modele charakteryzowane był przez następujące wartości błędu:
$CCR_{(KS)} = 95{,}83\%$;
$CCRT_{(KS)} = 100\%$
oraz
$CCR_{(DU)} = 97{,}92\%$;
$CCRT_{(DU)} = 100\%$.

Sieci neuronowe

Modele ANN oraz NFS, w przeciwieństwie do modeli CART oraz PLS, nie były konstruowane w oparciu o całe sygnały instrumentalne, ale w oparciu o dane zredukowane. Oryginalne dane zastąpiono czynnikami głównymi i zmiennymi istotnymi wskazanymi przez model CART. Przed analizą zmienne poddano skalowaniu do przedziału <-1, 1>.

Skonstruowany model ANN zawierał we wszystkich węzłach warstwy ukrytej oraz wyjściowej funkcję tangens hiperboliczny. Jako pierwszy modelowany zestaw danych użyto czterech czynników głównych (PCs) opisujących 98,57% wariancji danych. Optymalna sieć zawiera cztery węzły wejściowe i po jednym węźle w warstwie ukrytej oraz wyjściowej. Sieć ta pozwoliła na przewidzenie rodzaju dodatku paszowego dla zbiorów utworzonych za pomocą algorytmu Kennarda i Stone'a z następującymi błędami:
$CCR_{(KS/4PCs)} = 100\%$;
$CCRT_{(KS/4PCs)} = 100\%$.

Kolejny zestaw danych zawierał zmienne istotne (ZM: 1, 151, 154, 174) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Optymalny model to sieć zawierająca cztery węzły wejściowe i po jednym węźle w warstwie ukrytej oraz wyjściowej. Pozwoliła ona na przewidzenie modelowanej własności z następującym powodzeniem:
$CCR_{(KS/4ZM)} = 100\%$;
$CCRT_{(KS/4ZM)} = 100\%$.

Następny modelowany zestaw danych to cztery czynniki główne opisujące 98,57% wariancji danych zawierających obiekty podzielone na zbiory za pomocą algorytmu Duplex. Optymalna sieć zawierała cztery węzły wejściowe oraz po jednym węźle w warstwie ukrytej i wyjściowej. Moc predykcyjną tego modelu określały następujące wartości błędu:
$CCR_{(DU/4PCs)} = 100\%$;
$CCRT_{(DU/4PCs)} = 100\%$.

Ostatni zestaw danych zawierał zmienne istotne (ZM: 152, 179) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Optymalny model to sieć zawierająca dwa węzły wejściowe oraz po jednym węźle w warstwie ukrytej i wyjściowej. Pozwoliła ona na przewidzenie modelowanej własności z następującymi błędami:
$CCR_{(DU/2ZM)} = 100\%$;
$CCRT_{(DU/2ZM)} = 100\%$.

Neuronowe systemy rozmyte

Do modelowania danych zawierających zbiory uzyskane algorytmem Kennarda i Stone'a (KS) oraz algorytmem Duplex (DU) zastosowano modele NFS typu Sugeno pierwszego rzędu Jako pierwszy modelowany zestaw danych użyto czterech czynników

głównych (PCs) opisujących 98,57% wariancji danych. Obiekty podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Uzyskane wyniki były jednakowe bez względu na sposób iteracyjnego uczenia modelu. Skonstruowany model pozwolił na przewidzenie rodzaju dodatku paszowego z następującymi błędami:
$CCR_{(KS/4PCs)} = 100\%$;
$CCRT_{(KS/4PCs)} = 100\%$.

Kolejny modelowany zestaw danych zawierał zmienne istotne (ZM: 1, 151, 154, 174) wybrane przez model CART dla danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Skonstruowany model pozwolił na przewidzenie modelowanej własności z następującym sukcesem:
$CCR_{(KS/4ZM)} = 100\%$;
$CCRT_{(KS/4ZM)} = 100\%$.

Trzeci modelowany zestaw danych to cztery czynniki główne opisujące 98,57% wariancji danych, dla którego obiekty podzielono na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystywał schemat kratkowy do podziału przestrzeni danych, poprzez nałożenie dwóch funkcji przynależności na każdą ze zmiennych. W ramach tego modelu konstruowanych było szesnaście reguł logicznych. Uczenie modelu odbywało się w oparciu o wsteczną propagację błędu, a model pozwolił na przewidzenie modelowane własności z następującymi błędami:
$CCR_{(DU/4PCs)} = 97{,}92\%$;
$CCRT_{(DU/4PCs)} = 100\%$.

Czwarty zestaw danych zawierał zmienne istotne (ZM: 152, 179) wybrane przez model CART dla danych zawierających zbiory otrzymane metodą Duplex. Iteracyjne uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model do podziału przestrzeni danych wykorzystywał schemat kratkowy przypisując każdemu parametrowi po dwie funkcje przynależności. W ramach tego modelu skonstruowano cztery reguły logiczne. Model obarczony był błędami:
$CCR_{(DU/2ZM)} = 100\%$;
$CCRT_{(DU/2ZM)} = 100\%$.

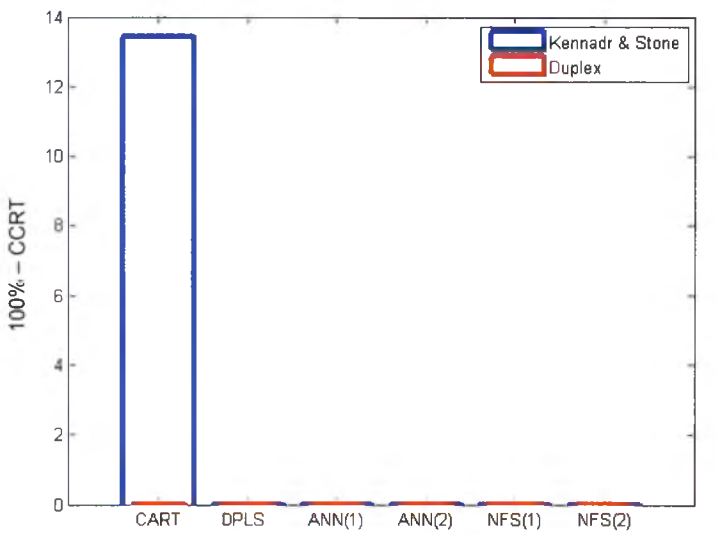
Podsumowanie

W tabeli 10 zamieszczono wyniki modelowania składu paszy zwierzęcej w oparciu o widma w bliskiej podczerwieni. Jakość modelu opisana jest za pomocą CCR i CCRT. Modele skonstruowano w oparciu o oryginalne zmienne (CART i PLS), lub czynniki główne i wybrane zmienne istotne (ANN, NFS).

Tabela 10 Zestawienie wyników przeprowadzonych analiz dla modelowania składu paszy zwierzęcej (Dane 8), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	CCR [%]	CCRT [%]	opis modelu
CART	KS	oryginalne	74,36	100	2 węzły terminalne
	DU	oryginalne	100	100	3 węzły terminalne
DPLS	KS	oryginalne	95,83	100	4 czynniki ukryte
	DU	oryginalne	97,92	100	4 czynniki ukryte
ANN	KS	4 PCs	100	100	1 węzeł w warstwie ukrytej
		4 ZM	100	100	1 węzeł w warstwie ukrytej
	DU	4 PCs	100	100	1 węzeł w warstwie ukrytej
		2 ZM	100	100	1 węzeł w warstwie ukrytej
NFS	KS	4 PCs	100	100	2 reguły logiczne
		4 ZM	100	100	2 reguły logiczne
	DU	4 PCs	97,92	100	16 reguł logicznych
		2 ZM	100	100	4 reguły logiczne

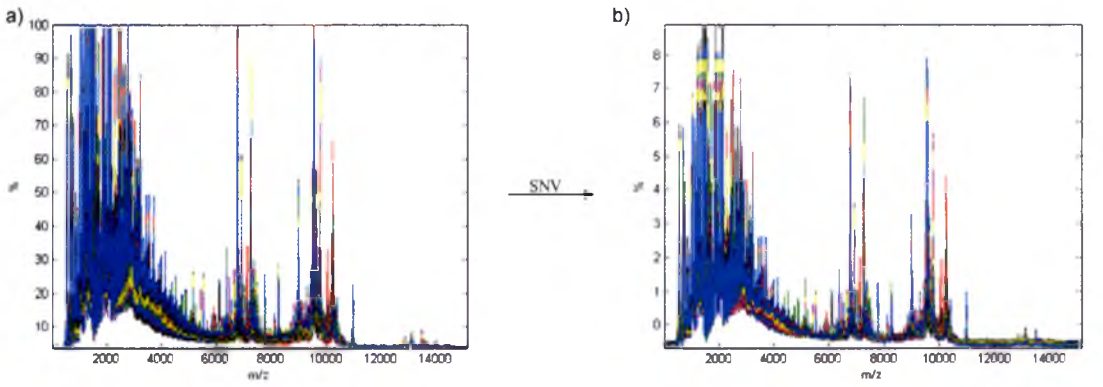
Procentowy wykres błędnie sklasyfikowanych próbek za pomocą zastosowanych metod przedstawiono na rysunku 106. Otrzymano wyższe wartości błędów dla modelu CART od wyników dla pozostałych metod modelowania danych. Moc predykcyjna modelu NFS była porównywalna do mocy predykcyjnej modeli DPLS i ANN. Co więcej metoda NFS ułatwia interpretację modelu dzięki regułom logicznym.



Rys. 106 Wykres procentu błędnie sklasyfikowanych próbek (100% – CCR) charakteryzujący konstruowane modele, gdzie indeksy oznaczają modele konstruowane w oparciu odpowiednio o dane zawierające (1) czynniki główne oraz (2) zmienne istotne

9.9 Dane 9: Modelowanie stanu zdrowia pacjentek

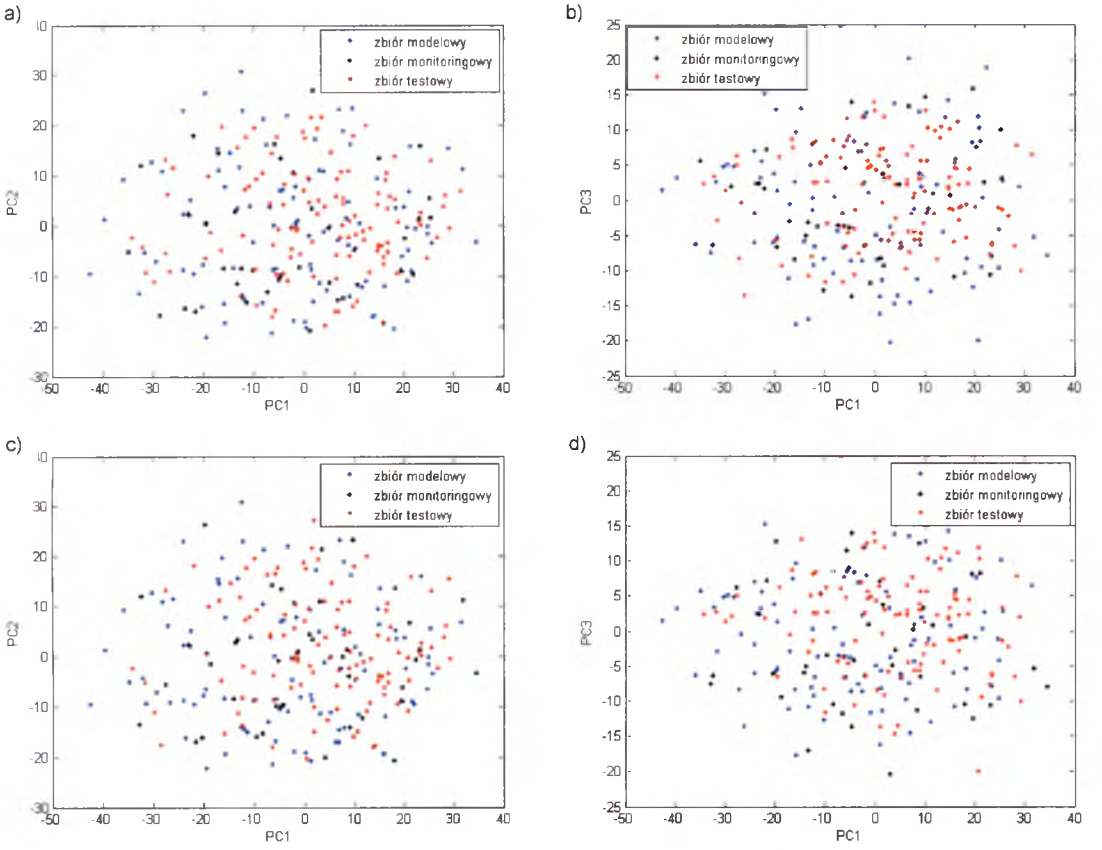
W ramach klinicznego projektu badawczego *Proteomics Databank* [113] badano kobiety cierpiące na nowotwór jajników. Zarejestrowano profile białkowe próbek krwi 253 pacjentek techniką spektroskopii mas niskiej rozdzielczości SELDI-TOF (Rys. 107a). Wśród badanych kobiet 162 cierpiały na nowotwór. Grupa kontrolna liczyła 91 osób. Dane o wymiarowości 253 x 15154 poddano wstępnej obróbce stosując transformacji SNV (Rys. 107b).



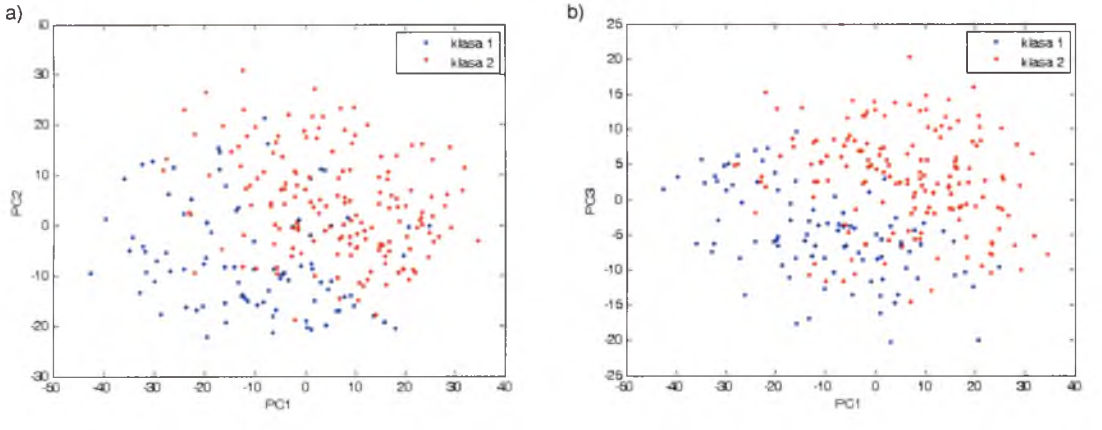
Rys. 107 Widma masowe 253 próbek krwi a) przed i b) po transformacji SNV

Eksploracja i przygotowanie danych

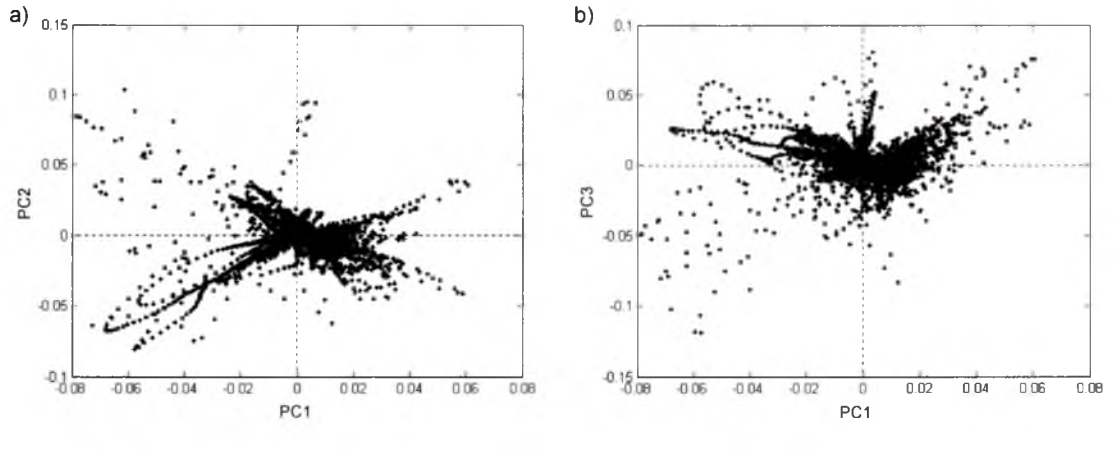
Zastosowano analizę czynników głównych (algorytm SVD2 [114]) celem wizualizacji i eksploracji danych (Rys. 108-111). W toku analizy nie wykryto obiektów odległych. Projekcje obiektów na płaszczyzny zdefiniowane przez wybrane czynniki główne (Rys. 109) uwidoczniły dwie grupy obiektów (klasa 1 i 2) nachodzących na siebie.



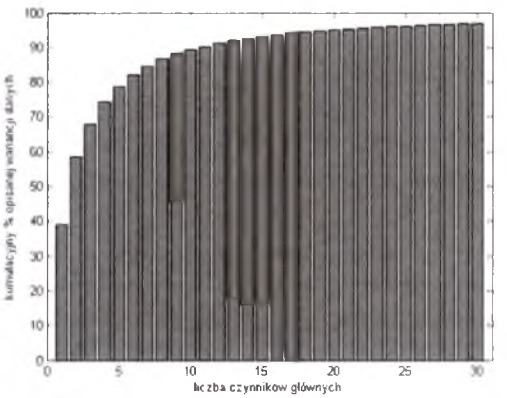
Rys. 108 Projekcja 253 obiektów na płaszczyznę zdefiniowaną przez: a, c) pierwszy i drugi czynnik główny oraz b, d) przez pierwszy i trzeci czynnik główny, gdzie wyboru obiektów do zbiorów dokonano za pomocą algorytmów: a, b) Kennarda i Stone'a i c, d) algorytmu Duplex



Rys. 109 Projekcja 253 obiektów na płaszczyznę zdefiniowaną przez: a) pierwszy i drugi czynnik główny oraz b) przez pierwszy i trzeci czynnik główny, gdzie zaznaczono stan zdrowia pacjentek (klasa 1 – osoby zdrowe, klasa 2 – osoby chore)



Rys. 110 Projekcja parametrów (m/z) na płaszczyznę wag zdefiniowana przez a) pierwszy i drugi czynnik główny oraz przez b) pierwszy i trzeci czynnik główny



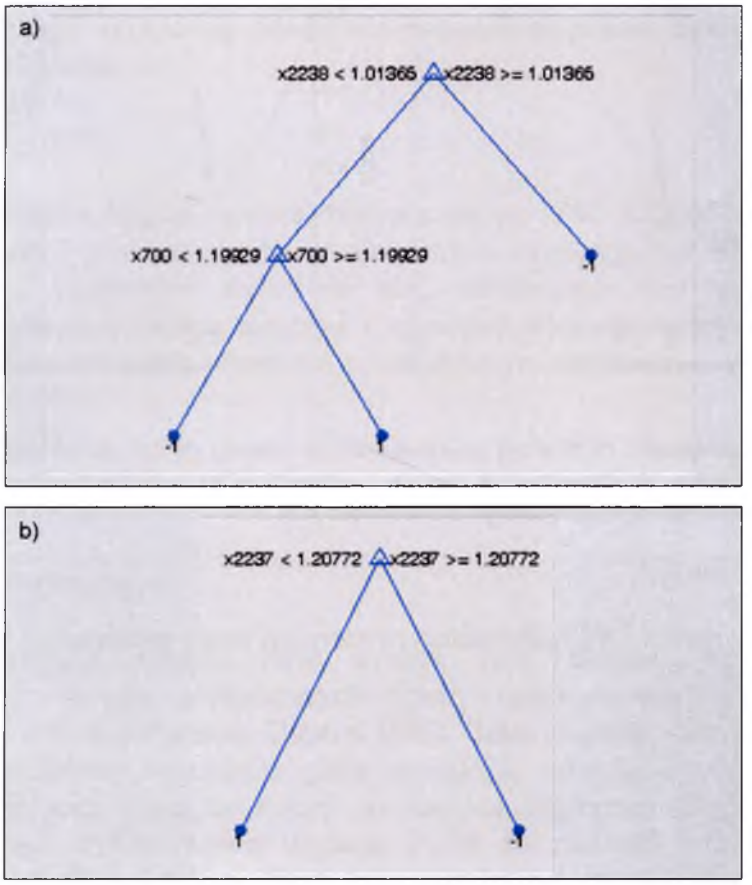
Rys. 111 Kumulacyjny procent opisanej wariancji danych przez kolejne czynniki główne

Następnie dane podzielono na trzy zbiory przypisując po 50 obiektów z każdej klasy do zbioru modelowego ($\mathbf{X}_{ml}$, $\mathbf{y}_{ml}$) oraz po 20 do zbioru monitoringowego ($\mathbf{X}_{mr}$, $\mathbf{y}_{mr}$) oraz resztę (21 obiektów z klasy 1/zdrowych oraz 92 z klasy 2/chorych) do zbioru testowego ($\mathbf{X}_{tt}$, $\mathbf{y}_{tt}$). Podziału na zbiory dokonano przy użyciu algorytmu Kennarda i Stone'a (KS) oraz algorytmu Duplex (DU). Zmienna zależna dla wszystkich zbiorów ($\mathbf{y}_{ml}$, $\mathbf{y}_{mr}$, $\mathbf{y}_{tt}$) została zakodowana binarnie. Tak utworzone zbiory zostały poddane analizie metodą CART oraz PLS.

Drzewa klasyfikacji i regresji

Następnie przystąpiono do konstrukcji modelu drzew klasyfikacji i regresji (CART). Optymalne binarne drzewo decyzyjne skonstruowane w oparciu o zbiory tworzone za pomocą algorytmu Kennarda i Stone'a miało trzy węzły terminalne (Rys. 112a). Zmienne wskazane w modelu jako decyzyjne to zmienne 700, 2238

oraz zmienna 485 wskazana przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły odpowiednio:
$CCR_{(KS)} = 95{,}24\%$;
$CCRT_{(KS)} = 98{,}23\%$.



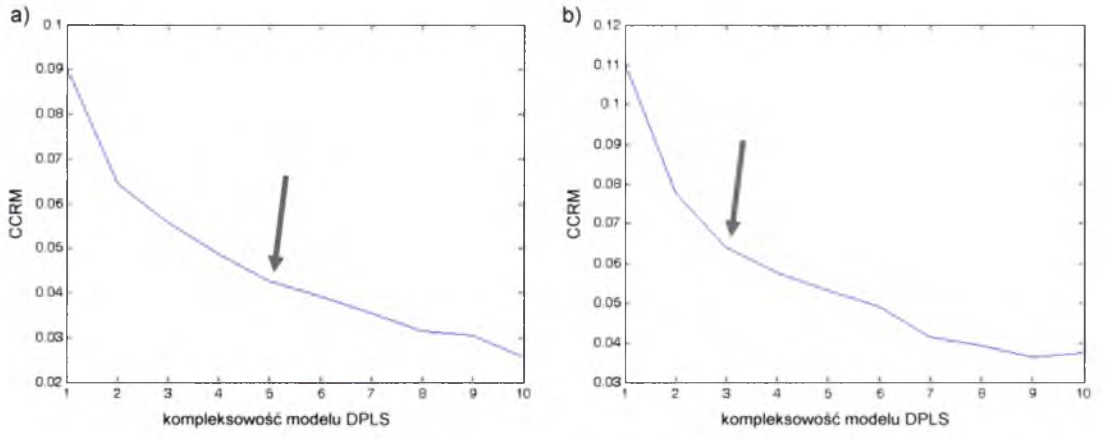
Rys. 112 Optymalne drzewo CART skonstruowane celem klasyfikacji pacjentek zdrowych i chorych na nowotwór jajników w oparciu o zbiory utworzone za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie (1) klasa 1 i (-1) klasa 2

Drzewo CART będące optymalnym modelem, skonstruowane w oparciu o dane zawierające zbiory utworzone za pomocą algorytmu Duplex, miało dwa węzły terminalne (Rys. 112b). Zmienne wskazane w modelu jako decyzyjne to zmienna 2237 oraz zmienne 850, 452 wskazane przez model przed przycinaniem drzewa. Wartości błędów dla tego modelu wyniosły:
$CCR_{(DU)} = 85{,}71\%$;
$CCRT_{(DU)} = 96{,}46\%$.

Metoda częściowych najmniejszych kwadratów w wariancie dyskryminacyjnym

Skonstruowano model DPLS w oparciu o dane zawierające zbiory otrzymane za pomocą algorytmu Kennarda i Stone'a (KS, Rys. 113a) i algorytmu Duplex

(DU, Rys. 113b). Optymalna kompleksowość modeli DPLS to odpowiednio pięć czynników ukrytych (KS) oraz trzy czynniki ukryte (DU).



Rys. 113 Wykres zależności CCRM od kompleksowości modelu DPLS dla zbioru monitoringowego utworzonego za pomocą a) algorytmu Kennarda i Stone'a (KS) oraz b) algorytmu Duplex (DU), gdzie strzałką zaznaczono optymalną kompleksowość modelu

Końcowy model DPLS charakteryzowany był przez następujące wartości błędu:
$CCR_{(KS)} = 100\%$;
$CCRT_{(KS)} = 100\%$
oraz
$CCR_{(DU)} = 100\%$;
$CCRT_{(DU)} = 100\%$.

Sieci neuronowe

Do konstrukcji modeli ANN oraz NFS wykorzystywano czynniki główne, a także zmienne wybrane przez model CART. Nowe zmienne poddano skalowaniu do przedziału <-1, 1>. Konstruowane modele ANN zawierały w węzłach warstwy ukrytej oraz w węźle warstwy wyjściowej funkcję typu tangens hiperboliczny. Jako pierwszy modelowany zestaw danych użyto jedenastu czynników głównych (PCs) opisujących 90,34% wariancji danych. Obiekty podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalna sieć zawierała jedenaście węzłów wejściowych i po jednym węźle w warstwie ukrytej oraz wyjściowej. Sieć ta pozwoliła na przewidzenie stanu pacjentki z następującymi błędami:
$CCR_{(KS/11PCs)} = 100\%$;
$CCRT_{(KS/11PCs)} = 100\%$.

Kolejny zestaw danych zawierał zmienne istotne (ZM: 485, 700, 2238) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Optymalny model to sieć zawierająca trzy węzły wejściowe, cztery węzły w warstwie ukrytej oraz jeden w warstwie wyjściowej. Sieć ANN pozwoliła na przewidzenie modelowanej własności z następującymi błędami:

$CCR_{(KS/3ZM)} = 100\%$;
$CCRT_{(KS/3ZM)} = 100\%$.

Następny modelowany zestaw danych to jedenaście czynników głównych (PCs) opisujących 90,34% wariancji danych podzielone na zbiory za pomocą algorytmu Duplex. Optymalna sieć zawiera jedenaście węzłów wejściowych oraz po jednym węźle w warstwie ukrytej i wyjściowej. Model ten pozwolił na przewidzenie stanu pacjentki z następującymi błędami:
$CCR_{(DU/11PCs)} = 100\%$;
$CCRT_{(DU/11PCs)} = 100\%$.

Ostatni zestaw danych zawierał zmienne istotne (ZM: 452, 856, 2237) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Optymalny model to sieć zawierająca trzy węzły wejściowe i po jednym węźle w warstwie ukrytej i wyjściowej. Skonstruowany model pozwolił na przewidzenie modelowanej własności z następującymi błędami:
$CCR_{(DU/3ZM)} = 98{,}00\%$;
$CCRT_{(DU/3ZM)} = 100\%$.

Neuronowe systemy rozmyte

Skonstruowano modele NFS według typu Sugeno pierwszego rzędu do modelowania danych zawierających zbiory uzyskane algorytmem Kennarda i Stone'a (KS) oraz algorytmem Duplex (DU). Jako pierwszy modelowany zestaw danych użyto jedenastu czynników głównych (PCs) opisujących 90,34% wariancji danych. Obiekty podzielono na zbiory za pomocą algorytmu Kennarda i Stone'a. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Uzyskane wyniki były jednakowe bez względu na sposób iteracyjnego uczenia modelu (wsteczna propagacja błędu lub metoda hybrydowa) Model NFS pozwolił na przewidzenie stanu zdrowia pacjentki z następującym powodzeniem:
$CCR_{(KS/11PCs)} = 100\%$;
$CCRT_{(KS/11PCs)} = 100\%$.

Drugi zestaw danych zawierał zmienne istotne (ZM: 485, 700, 2238) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Kennarda i Stone'a. Uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model wykorzystywał metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano 32 reguły logiczne. Skonstruowany model obarczony był błędami:
$CCR_{(KS/3ZM)} = 100\%$;
$CCRT_{(KS/3ZM)} = 100\%$.

Następny modelowany zestaw danych to jedenaście czynników głównych (PCs) opisujące 90,34% wariancji danych zawierających obiekty podzielone na zbiory za pomocą algorytmu Duplex. Optymalny model wykorzystuje metodę FCM do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Uzyskane wyniki były jednakowe bez względu na sposób iteracyjnego

uczenia modelu. Skonstruowany model pozwolił na przewidzenie stanu zdrowia pacjentki z następującymi błędami:
$CCR_{(DU/11PCs)} = 100\%$;
$CCRT_{(DU/11PCs)} = 100\%$.

Czwarty zestaw danych zawierał zmienne istotne (ZM: 452, 856, 2237) wybrane przez model CART podczas modelowania danych zawierających zbiory otrzymane metodą Duplex. Iteracyjne uczenie modelu odbywało się w oparciu o metodę hybrydową. Optymalny model wykorzystywał metodę grupowania różnicowego (o promieniu 1,0) do podziału przestrzeni danych. W ramach tego modelu skonstruowano dwie reguły logiczne. Model obarczony był błędami:
$CCR_{(DU/3ZM)} = 99{,}00\%$;
$CCRT_{(DU/3ZM)} = 100\%$.

Podsumowanie

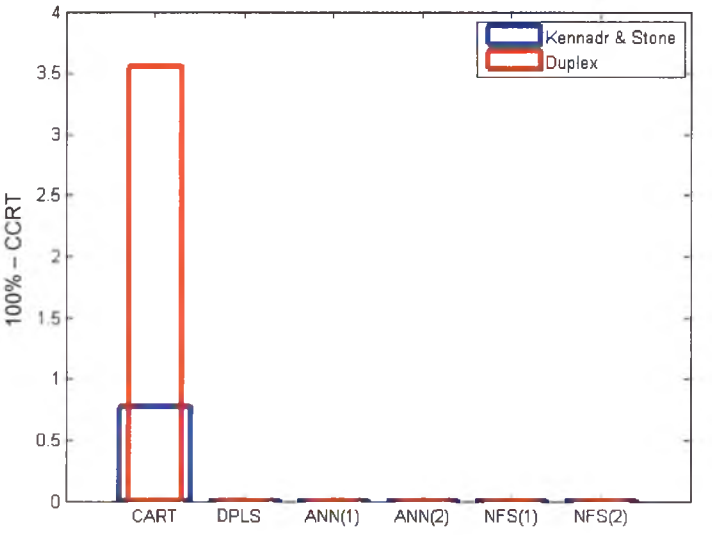
W tabeli 11 zamieszczono wyniki modelowania stanu zdrowia pacjentek w oparciu o widma masowe. Parametry charakteryzujące jakość skonstruowanych modeli czyli CCR i CCRT zamieszczono w czwartej i piątej kolumnie (Tabela 11). Z uwagi na wymiarowość danych modele ANN i NFS konstruowano w oparciu o czynniki główne i wybrane zmienne istotne. Modele CART i PLS były konstruowane w oparciu o oryginalne zmienne.

Tabela 11 Zestawienie wyników przeprowadzonych analiz dla modelowania stanu zdrowia pacjentek (Dane 9), gdzie KS i DU to odpowiednio skróty nazw algorytmu Kennarda i Stone'a oraz algorytmu Duplex

model	algorytm tworzenia zbiorów	modelowane zmienne	CCR [%]	CCRT [%]	opis modelu
CART	KS	oryginalne	95,24	98,23	3 węzły terminalne
	DU	oryginalne	85,71	96,41	2 węzły terminalne
DPLS	KS	oryginalne	100	100	5 czynników ukrytych
	DU	oryginalne	100	100	3 czynniki ukryte
ANN	KS	11 PCs	100	100	1 węzeł w warstwie ukrytej
		3 ZM	100	100	4 węzły w warstwie ukrytej
	DU	11 PCs	100	100	1 węzeł w warstwie ukrytej
		3 ZM	98,00	100	1 węzeł w warstwie ukrytej
NFS	KS	11 PCs	100	100	2 reguły logiczne
		3 ZM	100	100	32 reguły logiczne
	DU	11 PCs	100	100	2 reguły logiczne
		3 ZM	99,00	100	2 reguły logiczne

Na Rys. 114 przedstawiono procentowy wykres błędnie sklasyfikowanych próbek za pomocą zastosowanych metod. Wyniki dla modelu CART były mniej

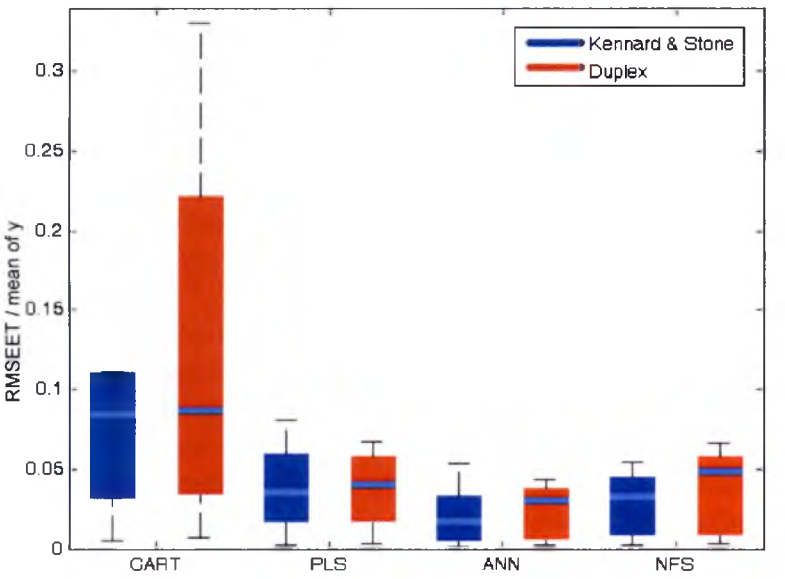
satysfakcjonujące od wyników dla pozostałych metod modelowania danych. Technika NFS pozwoliła na konstrukcję porównywalnych modeli do metod DPLS i ANN jednocześnie potencjalnie ułatwiając interpretację modelu dzięki regułom logicznym.



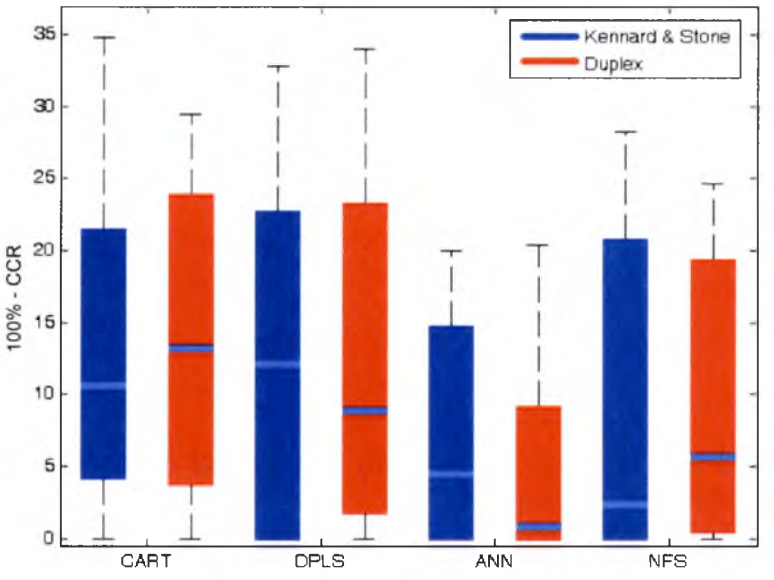
Rys. 114 Wykres procentu błędnie sklasyfikowanych próbek (100% – CCR) charakteryzujący konstruowane modele, gdzie indeksy oznaczają modele konstruowane w oparciu odpowiednio o dane zawierające (1) czynniki główne oraz (2) zmienne istotne

10 Podsumowanie

W niniejszej pracy podjęto próbę oceny możliwości zastosowań i efektywności neuronowych systemów rozmytych do modelowania danych chemicznych o zróżnicowanej strukturze. Jako parametr charakteryzujący jakość konstruowanych modeli przyjęto ich moc predykcyjną wyrażoną poprzez wartości błędów RMSEP (kalibracja) i CCRT (dyskryminacja). Poniżej zamieszczono wykresy charakteryzujące moc predykcyjną skonstruowanych modeli kalibracyjnych (Rys. 115) oraz klasyfikacyjnych (Rys. 116).



Rys. 115 Wykres charakteryzujący modele kalibracyjne – pierwiastek średniego błędu kwadratowego dla próbek z niezależnego zbioru testowego, przerywaną linią zaznaczono zakres otrzymanych wartości błędów, wypełnione słupki obejmują drugi i trzeci kwantyl wartości błędów, a fioletowa kreska obrazuje wartość wskazaną przez medianę



Rys. 116 Wykres charakteryzujący modele dyskryminacyjne – procent błędnie sklasyfikowanych próbek z niezależnego zbioru testowego; przerywaną linią zaznaczono zakres otrzymanych wartości błędów, wypełnione słupki obejmują drugi i trzeci kwantyl wartości błędów, a fioletowa kreska obrazuje wartość wskazaną przez medianę

Podsumowania przeprowadzonych analiz dokonano pod kątem kilku zagadnień. Po pierwsze porównano metodę NFS z trzema wybranymi technikami modelowania danych. Techniki te różniły się sposobem przetwarzania informacji (modele liniowe i nieliniowe) oraz charakterem konstruowanego modelu (lokalny i globalny). Porównano metodę NFS z metodą CART. Obie metody modelowania danych pozwalają na automatyczną konstrukcję reguł logicznych przez model w oparciu o analizowane dane. Jednak metoda CART nie zawsze pozwalała na konstrukcję modelu o satysfakcjonującej mocy predykcyjnej. Drugą wybraną metodą była metoda PLS, która jest powszechnie stosowaną techniką modelowania danych chemicznych z uwagi na prostotę modelu i łatwość w aplikacji samego algorytmu. Jest to jednak liniowa i globalna metoda modelowania danych co nie zawsze było pożądaną cechą konstruowanego modelu i czasem skutkowało pogorszeniem jego mocy predykcyjnej. Jako trzecią technikę modelowania danych, będącą odnośnikiem dla neuronowych systemów rozmytych, wybrano metodę ANN. Sieci neuronowe pozwalają na nieliniowe przetwarzanie danych co zaowocowało modelami o bardzo dobrej mocy predykcyjnej. Niestety modele ANN są bardzo trudne w interpretacji, w przeciwieństwie do modeli NFS.

Po drugie przedstawiono wyniki dla danych, gdzie modelowany problem miał charakter kalibracyjny, lub klasyfikacyjny. Analizowane dane różniły się liczebnością obiektów oraz parametrów, a także rozkładem obiektów w przestrzeni pomiarowej. Analizowano także dane po uprzedniej eliminacji obiektów odległych.

Na rysunku 115 widoczne jest, iż konstruowane modele kalibracyjne za pomocą neuronowych systemów rozmytych (NFS) były obarczone nieznacznie mniejszym błędem w porównaniu do metody PLS. Metoda NFS dawała także modele o lepszej mocy predykcyjnej w porównaniu do metody drzew klasyfikacji regresji CART.

Na uwagę zasługuje także fakt, iż tendencja ta nie była zależna od sposobu wyboru próbek do niezależnego zbioru testowego. Ponadto widoczne jest, iż sieci neuronowe pozwoliły na konstrukcję nieznacznie lepszych modeli w porównaniu do NFS. Należy jednak pamiętać, iż w przypadku sztucznych sieci neuronowych ewentualna interpretacja skonstruowanego modelu jest bardzo utrudniona. Rys. 116 charakteryzuje moc predykcyjną konstruowanych modeli dyskryminacyjnych. Podobnie jak w przypadku modeli kalibracyjnych, metoda NFS odznaczała się modelami o lepszej mocy predykcyjnej niż metoda CART oraz o porównywalnej mocy predykcyjnej do modeli ANN. Ponadto metoda NFS pozwala na konstrukcję modeli dyskryminacyjnych obarczonych mniejszym błędem niż metoda PLS.

Po trzecie porównano efektywność zastosowania metody NFS do modelowania skompresowanych danych zawierających czynniki główne z danymi zawierającymi wybrane zmienne istotne. Ponieważ istnieje wiele metod wyboru zmiennych istotnych, a przedmiotem niniejszej pracy nie było rozstrzyganie o wyższości którejkolwiek z nich, postanowiono wykorzystać zmienne decyzyjne z modelu CART jako zmienne istotne. Nie odnotowano bezpośredniej korelacji pomiędzy zastosowanym podejściem do redukcji wymiarowości danych, a mocą predykcyjną modelu NFS. Dylemat ten wymaga każdorazowo indywidualnego podejścia do analizowanych danych.

Po czwarte przeanalizowano możliwości zastosowań NFS w chemii i korzyści z tego płynących. Jak pokazano dzięki odpowiednim zabiegom transformacji danych metoda NFS nadaje się do modelowania danych chemicznych o szerokim spektrum pochodzenia nie ustępując innym metodą pod względem mocy predykcyjnej.

Wszystkie cztery aspekty oceny neuronowych systemów rozmytych przenikają się wzajemnie, a ostatecznym wyznacznikiem efektywności metody jest wartość jego mocy przewidywania modelowanej własności. W niniejszej pracy wykorzystywano dwie miary mocy predykcyjnej modeli RMSEP i CCR odpowiednio dla modeli kalibracyjnych i dyskryminacyjnych. Z uwagi na różny charakter opisywanych problemów obie miary błędów nie są porównywane miedzy sobą. Jednakże można w oparciu o otrzymane wyniki wyprowadzić uogólnione wnioski.

Na uwagę zasługuje także fakt, iż tendencja ta nie była zależna od sposobu wyboru próbek do niezależnego zbioru testowego. Ponadto widoczne jest, iż sieci neuronowe pozwoliły na konstrukcję nieznacznie lepszych modeli w porównaniu do NFS. Należy jednak pamiętać, iż w przypadku sztucznych sieci neuronowych ewentualna interpretacja skonstruowanego modelu jest bardzo utrudniona. Rys. 116 charakteryzuje moc predykcyjną konstruowanych modeli dyskryminacyjnych. Podobnie jak w przypadku modeli kalibracyjnych, metoda NFS odznaczała się modelami o lepszej mocy predykcyjnej niż metoda CART oraz o porównywalnej mocy predykcyjnej do modeli ANN. Ponadto metoda NFS pozwala na konstrukcję modeli dyskryminacyjnych obarczonych mniejszym błędem niż metoda PLS.

Po trzecie porównano efektywność zastosowania metody NFS do modelowania skompresowanych danych zawierających czynniki główne z danymi zawierającymi wybrane zmienne istotne. Ponieważ istnieje wiele metod wyboru zmiennych istotnych, a przedmiotem niniejszej pracy nie było rozstrzyganie o wyższości którejkolwiek z nich, postanowiono wykorzystać zmienne decyzyjne z modelu CART jako zmienne istotne. Nie odnotowano bezpośredniej korelacji pomiędzy zastosowanym podejściem do redukcji wymiarowości danych, a mocą predykcyjną modelu NFS. Dylemat ten wymaga każdorazowo indywidualnego podejścia do analizowanych danych.

Po czwarte przeanalizowano możliwości zastosowań NFS w chemii i korzyści z tego płynących. Jak pokazano dzięki odpowiednim zabiegom transformacji danych metoda NFS nadaje się do modelowania danych chemicznych o szerokim spektrum pochodzenia nie ustępując innym metodą pod względem mocy predykcyjnej.

Wszystkie cztery aspekty oceny neuronowych systemów rozmytych przenikają się wzajemnie, a ostatecznym wyznacznikiem efektywności metody jest wartość jego mocy przewidywania modelowanej własności. W niniejszej pracy wykorzystywano dwie miary mocy predykcyjnej modeli RMSEP i CCR odpowiednio dla modeli kalibracyjnych i dyskryminacyjnych. Z uwagi na różny charakter opisywanych problemów obie miary błędów nie są porównywane miedzy sobą. Jednakże można w oparciu o otrzymane wyniki wyprowadzić uogólnione wnioski.

11 Wnioski

Po przeanalizowaniu prezentowanych zestawów danych można konkludować, że:

- Neuronowe systemy rozmyte dają lepsze wyniki od drzew klasyfikacji i regresji jednocześnie dostarczając reguły logiczne.

- Wraz ze wzrostem liczby parametrów w danych i liczby stosowanych funkcji przynależności bardzo szybko rośnie liczba reguł logicznych, tworzonych w modelu NFS. Dlatego dane, które można efektywnie analizować za pomocą NFS powinny zawierać maksymalnie kilkanaście parametrów. Wielowymiarowe dane można poddać kompresji, oznacza to jednak utratę możliwości interpretacji modelu w świetle oryginalnych zmiennych.

- Istnieją dane, których analiza wymaga zastosowania lokalnego oraz nieliniowego modelu. W takiej sytuacji, pomimo utraty możliwości interpretacji modelu, neuronowe układy rozmyte pozwalają na konstrukcję modelu o większej mocy predykcyjnej w porównaniu do powszechnie stosowanych metod modelowania danych.

12 Bibliografia

[1] http://www.trace.eu.org
[2] www.fera.defra.gov.uk
[3] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and qualimetrics: part A, Elsevier, Amsterdam, Holandia 1997
[4] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and qualimetrics: part B, Elsevier, Amsterdam, Holandia 1998
[5] S. Osowski, Sieci neuronowe, Oficyna wydawnicza Politechniki Warszawskiej, Warszawa, Polska 1996
[6] D. Rutkowska, M. Piliński, L. Rutkowski, Sieci neuronowe, algorytmy genetyczne i systemy rozmyte, Wydawnictwo Naukowe PWN, Warszawa, Polska 1997
[7] J.-S. Jang, C.-T. Sun, E. Mizutani, Neuro Fuzzy and Soft Computing, Prentice-Hall, Inc., Englewood Cliffs, USA 1997
[8] C.-T. Lin, C.S.G. Lee, Neural Fuzzy systems, A neuro-fuzzy Synergism to intelligent Systems, Prentice Hall P T R, Upper Sadle River, NJ, USA 1996
[9] L. Rutkowski, Flexible neuro-fuzzy systems, Kluwer Academic Publishers, Norwell, USA 2004
[10] L.A. Zadeh, Inform. Control 8 (1965) 338
[11] C. Carlsson, R. Fuller, Optimization under if-then rules, Fuzzy Sets and Systems, 119 (2001) 111-120
[12] W.V. Leekwijck, E.E. Kerre, Defuzzification: criteria and classification, Fuzzy Sets and Systems 108 (1999) 159-178
[13] E.H. Mamdani, Applications of fuzzy algorithm for control a simple dynamic plant, Proc. IEEE 121 (1974) 1585–1588
[14] E.H. Mamdani, Advances in the linguistic synthesis of fuzzy controllers, Internat. J. Man Mach. Stud. 8 (1976) 669–678
[15] E.H. Mamdani, Application of fuzzy logic to approximate reasoning using linguistic systems, IEEE Trans. Comput. 26 (1977) 1182–1191
[16] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modelling and control, IEEE Trans. Systems Man Cybernet. 15 (1985) 116–132
[17] M. Sugeno, G.T. Kang, Structure identification of fuzzy model, Fuzzy Sets and Systems 28 (1988) 116-132
[18] Y. Tsukamoto, An approach to fuzzy reasoning method, in, Advances in Fuzzy Set Theory and Applications, M.M. Gupta, R.K. Ragade, R.R. Yaged, Nrth-Holland, Amsterdam, Holland 1979
[19] S.-H. Liao, Expert system methodologies and applications – a decade review from 1995 to 2004, Expert Systems with Applications 28 (2005) 93-103
[20] R. Babuska, Fuzzy Modelling for Control, Kluwer Academic Publishers, Boston, USA 1998
[21] S. Chiu, J.J. Cheng, Automatic rule generation of fuzzy rule base for robot arm posture selection, in: Proc. of NAFIPS Conf., San Antonio, Texas, USA 1994
[22] T. Martin, B. Majeed, B.-S. Lee, N. Clarke, A third-generation telecare system using fuzzy ambient intelligence, Studies in Computational Intelligence, 72 (2007) 155-175
[23] B.A. Sproule, C.A. Naranjo, I.B. Türksen, Fuzzy pharmacology: theory and applications, TRENDS In Pharmacological Sciences 23 (2002) 412-417
[24] Z. Zhang, H. Zhou, S. Liu, P.B. Harrington, An application of Takagi-Sugeno fuzzy system to the classification of cancer patients based on elemental contents in serum samples, Chemometrics and Intelligent Laboratory Systems 82 (2006) 294-299

[25] S.-M. Guo, C.-S. Lee, C.-Y. Hsu, An intelligent image agent based on soft-computing techniques for color image processing, Expert Systems with Applications 28 (2005) 483-494
[26] F. Hoeppner, Fuzzy shell clustering algorithms in image processing: fuzzy c-rectangular and 2-rectangular shells, IEEE Trans. Fuzzy Systems 5 (4) (1997) 599–613.
[27] B.E.A. Fisher, Fuzzy approaches to environmental decisions: application to air quality, Environmental Science and Policy 9 (2006) 22-31
[28] J. Horiuchi, K. Hiraga, Industrial application of fuzzy control to large-scale recombinant vitamin B2 production, J. Ferment. Bioeng. 87 (1999) 365-371
[29] M. Hosobuchi, F. Fukui, H. Matsukawa, T. Suzuki, H. Yoshikawa, Fuzzy control during microbial production of ML-236B, a precursor of pravastin sodium, J. Ferment. Technol. 76 (1993) 482-486
[30] A. Perendeci, S. Arslan, S.S. Celebi, A. Tanyolac, Prediction of effluent quality of an anaerobic treatment plant under unsteady state through ANFIS modelling with on-line input variables, Chemical Engineering Journal 145 (2008) 78–85
[31] H. Honda, T. Kobayashi, Fuzzy control of bioprocess, Journal of Bioscience and Bioengineering 89 (2000) 401-408
[32] P.M. Larsen, Industrial applications of fuzzy logic control, Int. J. Man-Machine Studies 12 (1980) 3-10
[33] J.M. Andújar, J.M. Bravo, Multivariable fuzzy control applied to the physical-chemical treatment facility of a Cellulose factory, Fuzzy Sets and Systems 150 (2005) 475-492
[34] T. Nakamura, T. Kuratani, Y. Morita, Fuzzy control: application to glutamic acid fermentation, Proc of IFAC Modeling and Control of Biotechnology Process (1985) 211-215
[35] K. Oishi, M. Tominaga, A. Kawato, S. Imayasu, S. Nanba, Application of fuzzy control theory to the sake brewing process,J. Ferment. Bioeng. 72 (1990) 115-121
[36] M.J. Soneira, R. Perez-Pueyo, S. Ruiz-Moreno, Raman spectra enhancement with fuzzy logic approach, Journal of Raman Spectroscopy 33 (2002) 599-603
[37] B. Yan, T.R. McJunkin, D.L. Stoner, J.R. Scott, Validation of fuzzy logic method for automated mass spectral classification for mineral imaging, Applied Surface Science 253 (2006) 2011-2017
[38] M. Otto, H. Bandemer, A fuzzy method for component identification and mixture evaluation in the ultraviolet spectral range, Analytical Chimica Acta 191 (1986) 193-204
[39] T. Arnold, M. Otto, W. Wegscheider, Interpretation system for automated wavelength dispersive X-ray fluorescence spectrometry, Talanta 47 (1994) 1169-1184
[40] M. Otto, Fuzzy theory. A promising tool for computerized chemistry, Analytica Chmica Acta 235 (1990) 169-175
[41] M. Otto, H. Bandemer, Pattern recognition based on fuzzy observations for spectroscopic quality control and chromatographic fingerprinting, Analytical Chimica Acta 184 (1986) 21-31
[42] B. Walczak, W. Wu, Fuzzy warping of chromatograms, Chemometrics and Intelligent Laboratory Systems 77 (2005) 173-180
[43] R.O. Duda, P.E. Hart, D.G. Stork, Pattern classification, John Wiley & Sons, Inc., Toronto, Canada 2001
[44] Y.H. Bang, C.K. Yoo, I.-B. Lee, Nonlinear PLS modelling with fuzzy inference system, Chemometrics and Intelligent Laboratory Systems 64 (2003) 137-155
[45] J.C. Bezdek, Plenum, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, USA 1981
[46] J.T. Tou, R.C. Gonzales, Pattern Recognition Principles, Addison-Wesley, Reading, MA, 1974
[47] D. Dembélé, P. Kostne, Fuzzy C-means method for clustering microarray data ,Bioinformatics 19 (2003) 973-980
[48] T. Rutkowski, Sieci neuronowe, Akademicka Oficyna Wydawnicza RM, Warszawa, Polska 1993
[49] F. Rosenblatt, The perceprton. A probabilistic model for information storage and organization in the brain, Psychology Review 65 (1958) 386-408
[50] F. Rosenblatt, Principles of neurodynamics: perceptrons and the theory of brain mechanics, Spartan, New Yor, USA 1962
[51] J. Hertz, A. Krogh, R.G. Palmer, Wstęp do teorii obliczeń neuronowych, Wydawnictwa Naukowo-Techniczne, Warszawa, Polska 1995
[52] J.A. Andersen, E. Rosenfeld, Neurocomputing – Foundations of Research, MIT Press, Cambridge, Mass, USA 1988
[53] W.J. Melssen, J.R.M. Smits, L.M.C. Buydens, G. Katema, Using artificial neural networks for solving chemical problems, Part II: Kohonen self-organising feature maps and Hopfield networks, Chemometrics and Intelligent Laboratory Systems 23 (1994) 267-291
[54] K. Hornik, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359-366

[55] Comprehensive Chemometrics, Chemical and Biochemical Data Analysis, red. S.D. Bron, R. Tauler, B. Walczak, Elsevier, Oxford, UK 2009
[56] C.G. Looney, Advanced in feedforward neural networks: Demystifying knowledge acquiring black boxes, IEE Transactions on Knowledge and Data Engineering, 8 (1996) 211-226
[57] B.J. Wythoff, Backpropagation neural networks, A tutorial, Chemometrics and Intelligent Laboratory Systems 18 (1993) 115-155
[58] J.R.M. Smits, W.J. Melssen, L.M.C. Buydens, G. Kateman, Tutorial: Using artificial neural networks for solving chemical problems, Part I. Multi-layer feed-forward networks, Chemometrics and Intelligent Laboratory Systems 22 (1994) 165-189
[59] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137
[60] M. Daszykowski, B. Walczak, D.L. Massart, Representative subset selection, Analityca Chimica Acta 468 (2002) 91-103
[61] R.D. Snee, Validation of regression models: methods and examples, Technometrics 19 (1977) 415-428
[62] Biocybernetyka i inżynieria biomedyczna 2000, red. M. Nałęcz, Tom 6, Akademicka Oficyna Wydawnicza Exit, Warszawa, Polska 2000
[63] F. Despagne, D.L. Massart, Neural networks in multivariate calibration, Analyst 123 (1998) 157-178
[64] P.J. Werbos, Beyond regression: new tools for prediction and analysis in the behavioural sciences, PhD dissertation, Harvard Univ., USA 1974
[65] W. Wua, L. Li, J. Yang, Y. Liu, A modified gradient-based neuro-fuzzy learning algorithm and its convergence, Information Sciences 180 (2010) 1630–1642
[66] J.A. Hartigan, M.A. Wong, A k-means clustering algorithm, Applied Statistics 28 (1979) 100-108
[67] K. Hammouda, F. Karray, A comparative study of data clustering techniques, Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Kanada N2L 3G1
[68] J.C. Bezdek, Fuzzy mathematics in pattern classification, PhD thesis, Applied Math. Center, Cornell University, Ithaca, USA 1973
[69] J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York, USA 1981
[70] I. Bergrt, B.-H. Mevik, T, Næs, New modifications and applications of fuzzy C-means methodology, Computational Statistics & Data Analysis 52 (2008) 2403-2418
[71] S.L. Chiu, Fuzzy model identification based in cluster estimation, Journal of Intelligent and Fuzzy Systems 2 (1994) 267-278
[72] M.F. Abbod, D.G. von Keyserlingk, D.A. Linkens, M. Mahfouf, Survey of utilization of fuzzy technology in medicine and healthcare, Fuzzy Sets and Systems 120 (2001) 331-349
[73] L. Tarassenko, S. Roberts, Supervised and unsupervised learning in radial basis function classifiers, IEE Processing -Vision, Image Signal Processing 141 (1994) 210-216
[74] C.-F. Chen, X. Feng, J. Szeto, Identification of critical genes in microarray experiments by a neuro-fuzzy approach, Computational Biology and Chemistry 30 (2006) 372-381
[75] A. Porwal, E.J.M. Carranza, M. Hale, A hybrid neuro-fuzzy model for mineral potential mapping, Mathematical geology 36 (2004) 803-826
[76] J. Kim, N. Kasabov, HyFIS: adaptive neuro-fuzzy inference systems and their application to nonlinear dynamical systems, Neural Networks 12 (1999) 1301-1319
[77] R.P. Paiva, A. Dourado, Interpretability and learning in neuro-fuzzy systems, Fuzzy Sets and Systems 147 (2004) 17-38
[78] www.scopus.com
[79] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chamica Acta 185 (1986) 1-17
[80] S. Wold, M. Siostrom, L. Ericksson, PLS-regression: a BASIC tool of Chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109-130
[81] D. Douroumis, L.J. Hadjileontiadis, A. Fahr, Adaptive neuro-fuzzy modelling of poorly soluble drug formulations, Pharmaceutical Research 23 (2006) 1157-1164
[82] B. Yea, T. Osaki, K. Sugahara, R. Konishi, Improvement of concentration-estimation algorithm for inflammable gases utilizing fuzzy rule-based neural networks, Sensors and Actuators B 56 (1999) 181-188
[83] A. Gulbag, F. Temurtas, A study of quantitative classification of binary gas mixture using neural networks and adaptative neuro-fuzzy inference systems, Sensors and Actuators B 115 (2006) 252-262
[84] A.I. Abdel-Rahman, G.J. Lim, A nonlinear partial least squares algorithm using quadratic fuzzy inference system, Journal of Chemometrics 23 (2009) 530-537
[85] Y.L. Loukas, Adaptive neuro-fuzzy inference system: an instant architecture-free predictor for improved QSAR studies, Journal of Medicinal Chemistry 44 (2001) 2772-2783

[86] E. Buyukbingol, A. Sisman, M. Akyildis, F.N. Alparslan, A. Adejare, Adaptive neuro-fuzzy inference system (ANFIS): A new approach to predictive modelling in QSAR applications: A study of neuro-fuzzy modelling of PCP-based NMDA receptor antagonists, Bioorganic & Medicinal Chemistry 15 (2007) 4265-4282
[87] M. Jalali-Heravi, P. Shahbazikkah, A. Ghadiri-Bidhendi, QSAR study of psychiatric drugs using classification and regression trees with adaptive neuro-fuzzy inference system, QSAR & Combinatorial Science 27 (2008) 729-739
[88] M. Jalali-Heravi, M. Asadollahi-Baboli, Quantitative structure-activity relationship study of serotonin (5-HT_7) receptor inhibitors using modified ant colony algorithm and adaptive neuro-fuzzy inference system (ANFIS), European Journal of Medicinal Chemistry 44 (2009) 1463-1470
[89] B. Walczak, W. Wegscheider, Calibration of non-linear analytical systems by a neuro fuzzy approach, Chemometrics and Intelligent Laboratory Systems 22 (1994) 199-207
[90] P.B. Harrington, B.W. Pack, FLIN: fuzzy linear interpolating network, Analytica Chimica Acta 277 (1993) 189-197
[91] P. Rearden, P.B. Harrington, J.J. Karnes, C.E. Bunker, Fuzzy rule-building expert system classification of fuel using solid-phase microextraction two-way gas chromatography differential mobility spectrometric data, Analytical Chemistry 79 (2007) 1485-1491
[92] X. Sun, C.M. Zimmermann, G.P. Jackson, C.E. Bunker, P.B. Harrington, Classification of jet fuels by fuzzy rule-building expert systems applied to three-way data by fast gas chromatography—fast scanning quadrupole ion trap mass spectrometry, Talanta 83 (2011) 1260-1268
[93] B. Walczak, E. Bauer-Wolf, W. Wegscheider, A neuro-fuzzy system for X-ray spectra interpretation, Mikrochimica Acta 113 (1994) 153-169
[94] H. Takahashi, K. Masuda, T. Ando, T. Kobazashi, H. Honda, Prognostic predictor with multiple fuzzy neural models using expression profiles from DNA microarray for metastates of breast cancer, Journal of Bioscience and Bioengineering 98 (2004) 193-199
[95] W.L. Tung, C. Quek, GenSo-FDSS: a neural-fuzzy decision support system for pediatric ALL cancer subtype identification using gene expression data, Artificial Intelligence in Medicine 33 (2005) 61-88
[96] W.-D. Yu, H.-W. Lin, A VaFALCON neuro-fuzzy system for mining of incomplete construction databases, Automation in Construction 15 (2006) 20-32
[97] M. Jalali-Heravi, A. Kyani, S. Afsari-Mamaghani, A. Ghadiri-Bidhendi, Quantative structure-retention relationship study of benzodiazepines using adaptive neuro fuzzy inference system as a feature selection method, QSAR & Combinatorial Science 27 (2008) 407-416
[98] J.F. de Canete, P. del Saz-Orozco, S. Gonzalez-Perez, Application of Adaptive Neurofuzzy Control Using Soft Sensors to Continuous Distillation Computer Aided Chemical Engineering 25 (2008) 465-470
[99] T. Azzouz, A. Puigdomenech, M. Aragay, R. Tauler, Comparison between different data pre-treatment methods In the analysis of forage samples using near-infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method, Analytica Chimica Acta 484 (2003) 121-134
[100] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37-52
[101] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, B. Walczak, Robust statistics In data analysis – A review basic concepts, Chemometrics and Intelligent Laboratory Systems 85 (2007) 203-219
[102] P.J. Rousseeuw, A.M. Leroy, Robust estimates, residuals and outlier detection, John Wiley & Sons Inc., New York, USA 1987
[103] P. Geladi, K. Esbensen, Regression on multivariate images: principal component regression for modelling, prediction and visual diagnostic tools, Journal of Chemometrics 5 (1991) 97-111
[104] I-C. Yeh, Modelling slump flow of concrete using second-order regressions and
artificial neural networks, Cement and Concrete Composites, 29 (2007) 474-480
[105] J.H. Kalivas, Two data sets of near infrared spectra, Chemometrics and Intelligent Laboratory Systems 37 (1997) 255-259
[106] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction: Foundations and Applications, Springer, Berlin, Niemcy 2006
[107] V. Centner, J. Verdu-Andres, B. Walczak, D. Joun-Rimbaud, F. Despagne, O. de Nord, A comparison of multivariate calibration techniques applied to experimental NIR data sets, Applied Spectroscopy, 54 (2000) 608-623
[108] D. Coomans, M. Jonckheer, D.L. Massart, I. Broeckaert, P. Blockx, The application of linear discriminant analysis in the diagnosis thyroid diseases, Analytica Chimica Acta 103 (1978) 409-415
[109] M. Barker, W. Rayens, Partia least squares for discrimination, Journal of Chemometrics 17 (2003) 166-173

[110] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236

[111] M.M. Krishna Reddy, Priyankar Ghosh, S.N. Rasool, R.K. Sarin, R.B. Sashidhar, Source identification of Indian opium based on chromatographic fingerprinting of amino AIDS, Journal of Chromatography A, 1088 (2005) 158–168

[112] Dane dzięki uprzejmości dra Henryka Czarnik-Matusewicza z Katedry i Zakładu Farmakologii Klinicznej Akademii Medycznej we Wrocławiu.

[113] O.P. Whelehan, M.E. Earll, E. Johansson, M. Toft, L. Eriksson, Detection of ovarian cancer using chemometric analysis of proteomic profiles, Chemometrics and Intelligent Laboratory Systems 84 (2006) 82–87

[114] N. Kettaneh, A. Berglund, S. Wold, PCA and PLS with very large data sets, Computational Statistics & Data Analysis 48 (2005) 69-85

7. Dorobek naukowy

Publikacje:

- A.J. Charlton, M.S. Wróbel, I. Stanimirova, M. Daszykowski, H. Grundy, B. Walczak, Multivariate discrimination of wines with respect to their grape varieties and vintages, European Food Research and Technology, 231 (2010) 733-743
- M. Daszykowski, M.S. Wróbel, A. Bierczynska-Krzysik, J. Silberring, G. Lubec, B. Walczak, Automatic preprocessing of electrophoretic images, Chemometrics and Intelligent Laboratory Systems, 97 (2009) 132-140
- M. Daszykowski, M.S. Wróbel, H. Czarnik-Matusewicz, B. Walczak, Near-infrared reflectance spectroscopy and multivariate calibration techniques applied to model the protein, fiber and fat contents in rapeseed meal, The Analyst, 133 (2008) 1523-1531

Komunikaty ustne i postery:

- M. Sajewicz, D. Staszek, M. Wróbel, Ewa Recmanik, Kinga Sobczyk,Monika Waksmundzka-Hajnos, Teresa Kowalska, The HPLC analysis of the selected extract fractions derived from a variety of the sage (Salvia) species, conference on "Chromatographic methods of investigating the organic compounds", Szczyrk (9 czerwiec 2011), popster
- M.S. Wróbel, B. Walczak, The authentication of wine and olive oil samples in the European Union, II Międzynarodowe Sympozjum Edukacyjne: "Społeczeństwo i Zdrowie", Zakopane (24 listopad 2010), poster
- M.S. Wróbel, B. Walczak, Constructing logical rules for food tracebility, The 12th International Conference on Chemometrics in Analytical Chemistry, Antwerpia, Belgia (21 październik 2010), poster
- M.S. Wróbel, B. Walczak, Logical rules as a key for model interpretation, VIII Polska Konferencja Chemii Analitycznej *„Analityka dla Społeczeństwa* XXI wieku", Kraków (06 lipiec 2010), poster
- M.S. Wróbel, L. Komsta, M. Daszykowski, B. Walczak, Chemometrics in comparative analysis of samples, conference on "New analytical techniques in determining the quality of drugs", Lublin (25 wrzesień 2009), poster

- M.S. Wróbel, L. Komsta, M. Daszykowski, B. Walczak, Comparative analysis of chromatographic data, conference on "Chromatographic methods of investigating the organic compounds", Szczyrk (4 czerwiec 2009), poster
- M.S. Wróbel, M. Daszykowski, B. Walczak, Eliminacja zmiennych nieistotnych, a wieloparametrowa kalibracja, 4. Konferencja „Chemometria – Metody i Zastosowania", Zakopane (25 październik 2008), poster
- M.S. Wróbel, M. Daszykowski, B. Walczak, Pozyskiwanie istotnej informacji dla modelowania określonej własności, Konferencja „Chemometria – Metody i Zastosowania", Zakopane (25 październik 2008), komunikat ustny
- H. Czarnik-Matusewicz, M.S. Wróbel, M. Daszykowski, B. Walczak, Kontrola jakości śruty rzepakowej stosując spektroskopię w bliskiej podczerwieni oraz metody chemometryczne, 4. Konferencja „Chemometria – Metody i Zastosowania", Zakopane (24 październik 2008), poster
- M.S. Wróbel, M. Daszykowski, H. Czarnik-Matusewicz, B. Walczak, Controlling protein, fat and fiber content in animal feed using near infrared spectroscopy, II Conference on "Actual problems in analytical chemistry", Institute of Chemistry, Katowice (30 maj 2008), poster
- M.S. Wróbel, M. Daszykowski, B. Walczak, Is this peak pure?, The XXXIst Symposium Chromatographic Methods of Investigating The Organic Compounds, Szczyrk (04 czerwiec -2007), poster